Combining Classifiers Using the Dempster-Shafer Theory of Evidence

by

Imran Naseem

A Thesis Presented to the
DEANSHIP OF GRADUATE STUDIES

In Partial Fulfillment of the Requirements
for the Degree

MASTER OF SCIENCE

IN

Electrical Engineering

KING FAHD UNIVERSITY
OF PETROLEUM AND MINERALS

Dhahran, Saudi Arabia

January 2005

Dedicated to

My Beloved Parents

Mr. & Mrs. M.Naseem Siddiqui

and

My Advisor

Dr. Mohamed Deriche

ACKNOWLEDGEMENTS

In the name of Allah, the Most Gracious and the Most Merciful

All praise and glory goes to Almighty Allah (Subhanahu Wa Ta'ala) who gave me the courage and patience to carry out this work. Peace and blessings of Allah be upon His last Prophet Muhammad (Sallulaho-Alaihe-Wassalam) and all his Sahaba (Razi-Allaho-Anhum) who devoted their lives towards the prosperity and spread of Islam.

First and foremost gratitude is due to the esteemed university, the **King Fahd University of Petroleum and Minerals** for my admittance, and to its learned faculty members for imparting quality learning and knowledge with their valuable support and able guidance that has led my way through this point of undertaking my research work.

My deep appreciation and heartfelt gratitude goes to my thesis advisor **Dr. Mohamed Deriche** for his constant endeavour, guidance and the numerous moments of attention he devoted throughout the course of this research work. His valuable

suggestions made this work interesting and knowledgeable for me. Working with him in a friendly and motivating environment was really a joyful and learning experience.

I extend my deepest gratitude to my thesis committee members Dr. Asrar Sheikh and Dr. Mohandes.M for their constructive and positive criticism, extraordinary attention and thought-provoking contribution in my research. It was surely an honor and an exceptional learning to work with them.

Acknowledgement is due to my senior fellows Saad Azher and Mohammad Moin Uddin for helping me on issues relating to LATEX and MATLAB. I also appreciate the help provided by my fellow Mudassir Masood in programming on MATLAB.

Sincere friendship is the spice of life. I owe thanks to my house mates, colleagues and my friends for their help, motivation and pivotal support. A few of them are Moin Uddin, Saad Azhar, Sajid Khan, Mudassir Masood, Adnan Shahab, Khawar Khan, Kamran Arshad, Adnan Yousuf, Aiman Rasheed, Saad Mansoor, Imran Azam and many others; all of whom I will not be able to name here. They made my work and stay at KFUPM very pleasant and joyful. My heartfelt thanks to my days old friends Hashim, Farooq and Zeeshan . They truly are my great friends, I wish we could be together again.

Family support plays a vital role in the success of an individual. I would like to thank my parents, siblings, my grandfather and other family members including all my uncles, aunts and my loving cousins; from the core of my heart. Their prayers

and encouragement always help me take the right steps in life.

May Allah help us in following Islam according to Quran and Sunna! *(Aameen)*

Contents

List of Figures

List of Tables

Nomenclature

Abbreviations

ASI	Automatic Speaker Identification
ASR	Automatic Speaker Recognition
ASV	Automatic Speaker Verification
AWGN	Additive White Gaussian Noise
BBA	Basic Belief Assignment
bel	Belief
BKS	Behavior-Knowledge Space
BPA	Basic Probability Assignment
DCT	Discrete Cosine Transform
DST	Dempster-Shafer Theory of Evidence
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
ICA	Independent Component Analysis
LPCC	Linear Predictive Cepstral Coefficient

LDA	Linear Discriminant Analysis
MFCC	Mel Frequency Cepstral Coefficient
NN	Nearest Neighbor
NNEF	Nearest Neighbor based Evidence Fusion
PCA	Principal Component Analysis
rej	Rejection
RREF	Recognition Based Evidence Fusion
TBM	Transfer Belief Model
UPC	Universal Product Code
VEF	Variance Based Evidence Fusion

THESIS ABSTRACT

Name: Imran Naseem

Title: Combining Classifiers Using the Dempster Shafer Theory of Evidence

Degree: MASTER OF SCIENCE

Major Field: Electrical Engineering

Date of Degree: January 2005

As organizations strive for means of providing more secure methods for user access, biometrics is gaining increasing attention. However a biometric recognition system good for one case study may not be accurate for the other one. One solution to the problem is combining classifiers; so that the complementary information departed by different classifiers could be combined, in an efficient way, to achieve a much better recognition rate as compared to the participating experts. In this context the Dempster Shafer theory of evidence (DST) has shown some promising results; however the DST has not yet been explored for the problem of biometric recognition systems. In this thesis we have proposed three novel algorithms to combine different biometric systems using the DST. NNEF (Nearest Neighbor Based Evidence Fusion) algorithm uses the nearest neighbor distance of the participating experts as an evidence estimation parameter; RREF (Recognition Rate Based Evidence Fusion) algorithm uses the performance parameters of the participating experts for evidence estimation and VEF (Variance Based Evidence Fusion) algorithm uses the second order statistics of decision parameters to estimate the belief in the combining experts. Extensive experiments have been conducted on uni-modal (speech only) and multi-modal (speech and face) biometric recognition systems; the simulation results have shown that our proposed algorithms achieve much better recognition rate than the individual classifiers.

Keywords: *Pattern recognition, combining classifiers, biometrics, Dempster Shafer theory of evidence, multi-modal biometrics*

King Fahd University of Petroleum and Minerals, Dhahran.
January 2005

ملخص الرسالة

الاسم : **عمران نسيم**

العنوان: **دمج المصنفات باستخدام نظرية الدليل لدمبستر شافر**

الدرجة: **ماجستير**

قسم : **الهندسة الكهربائية**

التاريخ: **ذو القعدة ، 1426**

يزداد الإهتمام بالبيومتريكس بإزدياد إهتمام المنظمات في العالم الى التزود بوسائل أكثر ضمانا لحماية دخول المستخدمين لشبكاتها. على الرغم من أن نظام التمييز البيومتري قد يكون جيدا بالنسبة لحالة ما، قد لا يكون جيدا بالنسبة الى حالة أخرى. أحد الحلول لهذه المشكلة هو دمج المصنفات بحيث أنه يمكن دمج المعلومات المختلفة المستخلصة من مصنفات مختلفة، بفعالية عالية، من أجل الحصول على مستوى تمييز عالي مقارنة مع المصنفات لمشاركة. في هذا الشأن، إن نظرية الدليل للعالمين دمبستر وشافر(DST) قد أظهرت نتائج واعدة، على الرغم من أن ال(DST) لم تبحث من أجل مسائل التمييز البيومتري.

في هذه الرسالة، نحن اقترحنا ثلاث خوارزميات جديدية من أجل دمج الأنظمة البيومترية المختلفة باستخدام ال(DST). خوارزمية (NNFE) (دمج الدليل المعتمد على أقرب مجاور) تستخدم مسافة أقرب مجاور للمصنفات المشاركة كدليل لتقدير المتحول؛ خوارزمية (RRFE) (دمج الدليل المعتمد على معدل التمييز) تستخدم متحول الأداء للمصنفات المشاركة لتقدير الدليل وخوارزمية (VEF) (دمج الدليل المعتمد على التفاوت) تستخدم الإحصاء من الدرجة الثانية للمتحولات لتقدير المعرفة التخيلية في المصنفات المندمجة.

قد أجريت تجارب مكثفة على نموذج أحادي (الكلام فقط) ونموذج متعدد (الكلام والوجه) لأنظمة التمييز البيومترية؛ نتائج المحاكاة الحاسوبية أظهرت أن خوارزميتنا المقترحة تنجز معدل تمييز أفضل بكثير مما تنجزه المصنفات الأحادية.

هذه الدراسة اعدت لنيل درجة الماجستير في العلوم
في جامعة الملك فهد للبترول والمعادن
الظهران 31261

Chapter 1

Introduction

1.1 The Problem of Pattern Classification

Pattern classification (or recognition) is one of the key features of intelligent behavior for both humans and machines. It plays an important role in the daily life of humans, e.g. recognizing the faces of friends in a crowd, characters and words on printed pages, voices over a telephone line, and so forth. Likewise, reliance on machines that perform some sort of pattern recognition is increasing by the day. Examples of these machines are the readers of UPC (universal product code) bar codes that expedite pricing and inventory of retail merchandise, readers of magnetic-strip codes on credit cards that identify the user and determine if the purchase is authorized, etc.

Pattern recognition can be defined as the science that involves the description

or classification of measurements [3] .There is no doubt that pattern recognition is an important, useful, and rapidly developing technology with cross-disciplinary interest and participation. Some of the emerging applications of pattern recognition include: radar signal classification/analysis, image analysis, computer vision, face recognition, fingerprint identification, character recognition, speech recognition/ understanding, speaker identification, electroencephalogram (EEG) understanding/analysis, and medical diagnosis, etc. Human experts or even normal humans master most of these tasks quite easily. For many engineers and scientists, the ability to build machines which can perform such tasks as accurate as humans represents the ultimate challenge. Because of this, it has been of great interest to researchers to understand how humans process and analyze different incoming signals.

Human senses process signals, such as sounds or light waves, by transforming these in some way such that important information is extracted from these signals. The transformed signals are then mapped into a decision that equates with the recognition of objects. Such processing detects subtle differences in the signals that are necessary to perform optimal recognition. In order to design machines capable of classifying and recognizing patterns, observation vectors (e.g. collected from a probe or a camera) have to be transformed into feature vectors in a way similar to the processing of signals by humans. The features are intended to be fewer in number than the observations but should collectively contain most of the information

needed for classification of the patterns. This is simply because when the number of features is large, it becomes difficult to obtain good estimates for the parameters needed by the decision rule. For example, the number of pixels (picture elements) in a particular 1024X768 image is 786,432. Each 8-bit pixel may represent one of 256 shades of gray (or colors). Rather than using this huge number of observations, we could rely on a small number of important attributes. Some of the important features used for the classification of images include: angles between edges, and blobs of a particular shade. These features do not depend upon the size, location or orientation, but may be dilated, contracted, rotated, or translated. They may number only a few dozen or even less.

Algorithms that analyze data in an attempt to estimate appropriate features are called feature extraction algorithms. Such algorithms may be based on physical or structural considerations of the problem or they may be purely mathematical techniques. An example of physical sensory features used by machines is an image provided by a video camera. Structural features are relationships of physical sensory features, such as the relative locations of certain lines, edges, curves and blobs. A mathematical feature is obtained by mapping pattern observations via a function such that the newly obtained features have the power to distinguish between different categories or classes in a more efficient way.

The performance of the classifier is another important aspect that affects the overall behavior of pattern classification systems. The human brain can be consid-

ered as an optimal classifier. It can easily assign a certain pattern to a specific class label with a high degree of accuracy. However, this is not the case for machines. One way for improving the classification accuracy of machines is by testing more than one classification algorithm. The performance of these classifiers would then be assessed and a choice is made on the best performing classifier. However, detailed analysis of the performance of different classifiers showed that they tend to exhibit different misclassified patterns [4]. This means that if several classifiers are used to perform a specific classification task, then a wrong classification made by one classifier could be recovered, given that other classifiers are able to provide the correct answer. Thus, combining the classification results of different classifiers can help improve the performance of pattern classification systems.

1.2 Biometric Recognition Systems

One major application of pattern recognition systems is in person identification. With the exponential growth in technology and the growth in business carried worldwide, it is becoming crucial to build automated systems that identify people. Traditionally, body characteristics such as face and voice have been successfully used in identification. As early as the mid 19th century, Alphonse Bertillon, chief of the criminal identification division of the police department in Paris, actually developed the idea of using various body measurements [1] (for example, height, length of

arms, feet, and fingers) to identify criminals. In the late 19th century, just as his idea was gaining popularity, it was eclipsed by a far more significant and practical discovery: the distinctiveness of human fingerprints. Soon after this discovery, many major law-enforcement departments embraced the idea of "booking" criminals' fingerprints and storing them in databases (initially, card files). Later, police gained the ability to "lift" leftover, typically fragmentary, fingerprints from crime scenes (commonly called latents) and match these to fingerprints in the database to determine criminals identities. Biometrics first came into extensive use for law-enforcement and legal purposes, identification of criminals and illegal aliens. It was expanded for usage in security clearances for employees in sensitive jobs, paternity determinations, forensics, positive identifications of convicts and prisoners, and so on. Today, however, many civilian and private-sector applications are increasingly using biometrics to establish personal identification.

1.2.1 What is a Biometric?

Traditionally there have been three different types of person authentication:

- Something you know,a password, PIN, or piece of personal information
- Something you have, a card key, smart card, or token (like a SecurID card) and/or
- Something you are, a biometric.

From the list above, a biometric is the most secure and convenient authentication tool. It can't be borrowed, stolen, or forgotten, and forging one is practically impossible. Biometrics measure individuals' unique physical or behavioral characteristics to recognize or authenticate their identity. Common physical biometrics include: fingerprints, hand or palm geometry, retina, iris and facial characteristics. Behavioral characters include signature, voice (which also has a physical component), keystroke pattern, and gait. We discuss the main biometrics used nowadays in the next section.

1.3 Different Biometrics

1.3.1 Fingerprints

Figure 1.1: A typical fingerprint image (*www.cse.ucsd.edu*)

A fingerprint image looks at the patterns found on a fingertip (see figure 1.1). There are a variety of approaches to fingerprint verification. Some emulate the traditional police method of matching minutiae; others use straight pattern-matching devices; and still others are a bit more unique. Fingerprint identification systems rely on a set of features representing the ridge endings and bifurcations for matching patterns. Some verification approaches can detect when a live finger is presented; some cannot.

1.3.2 Hand Geometry

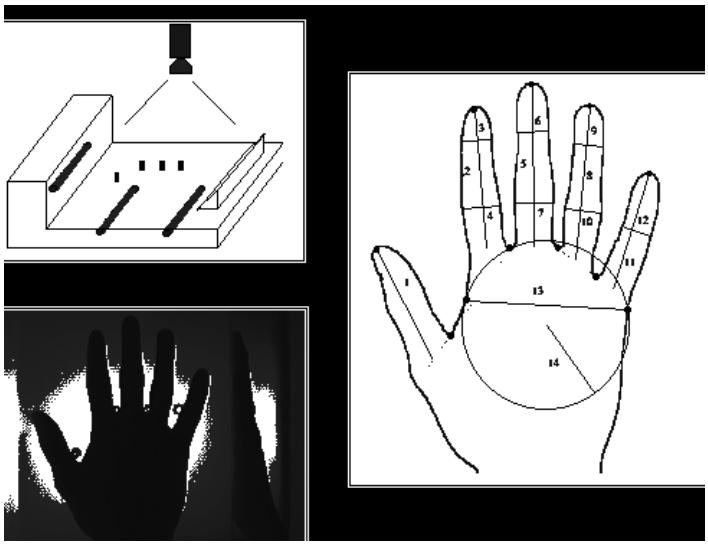

Figure 1.2: A typical human hand recognition system (*http://bias.csr.unibo.it*)

Hand geometry involves analyzing and measuring the shape of the hand (see figure 1.2).This biometric offers a good balance of performance characteristics and is relatively easy to use. It might be suitable where there are more users or where users

access the system infrequently and are perhaps less disciplined in their approach to the system. Hand recognition systems use mainly geometric features such as width of fingers, palm etc. Accuracy can be very high if desired, and flexible performance tuning and configuration can accommodate a wide range of applications.

1.3.3 Retina and Iris

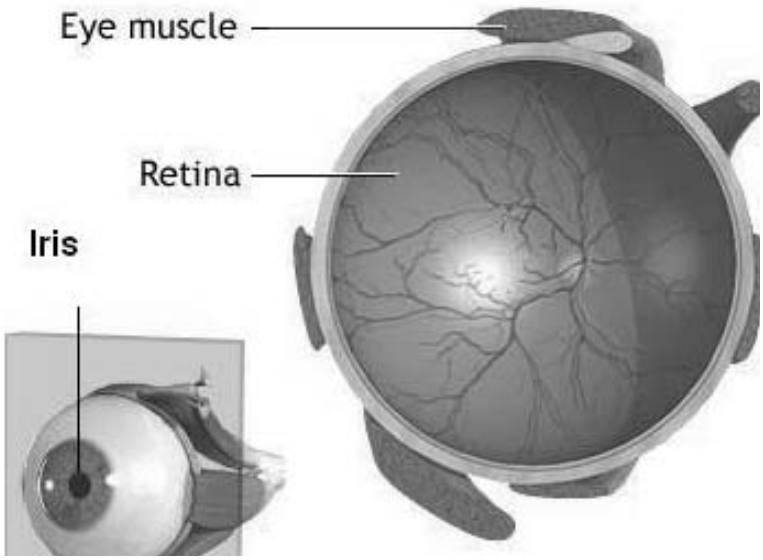

Figure 1.3: The human eye image (*www.nlm.nih.gov*)

A retina-based biometric involves the analysis of the layer of blood vessels at the back of the eye (see figure 1.3). An established technology, this technique involves using a low intensity light source through an optical coupler to scan the unique patterns of the retina. Retinal scanning can be quite accurate but does require the user to look into a receptacle and focus on a given point.This is not particularly convenient if you wear glasses or are concerned about having close contact with the

reading device. For these reasons, retinal scanning is not warmly accepted by all users, even though the technology itself can work well.

An iris-based biometric, on the other hand, involves analyzing features found in the colored ring of tissue that surrounds the pupil. Iris scanning, undoubtedly the less intrusive of the eye related biometrics, uses a fairly conventional camera element and requires no close contact between the user and the reader. In addition, it has the potential for higher than average template- matching performance. Iris biometrics work with glasses in place and is one of the few approaches that can work well in identification mode.

1.3.4 Face

Figure 1.4: A typical human face image (*www.uk.research.att.com*)

Face recognition systems analyze facial characteristics (see figure 1.4). A digital

camera is required to capture a facial image of the user for authentication. This technique has attracted considerable interest. Details about face recognition are discussed later in the thesis.

1.3.5 Signature

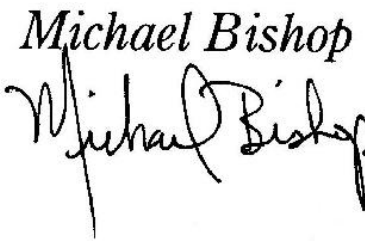

Figure 1.5: A typical signature (*http://bellsouthpwp.net*)

Signature verification analyzes the way a user signs his/her name (see figure 1.5). Signing features such as speed, velocity, and pressure are as important as the finished signatures static shape. Signature verification enjoys a synergy with existing processes that other biometrics do not. People are used to signatures as a means of transaction-related identity verification, and most would see nothing unusual in extending this to encompass biometrics. Surprisingly, relatively few significant automated signature applications have emerged compared with other biometric methodologies. This is mainly due to the changes of the signature of the same individual over time.

1.3.6 Voice

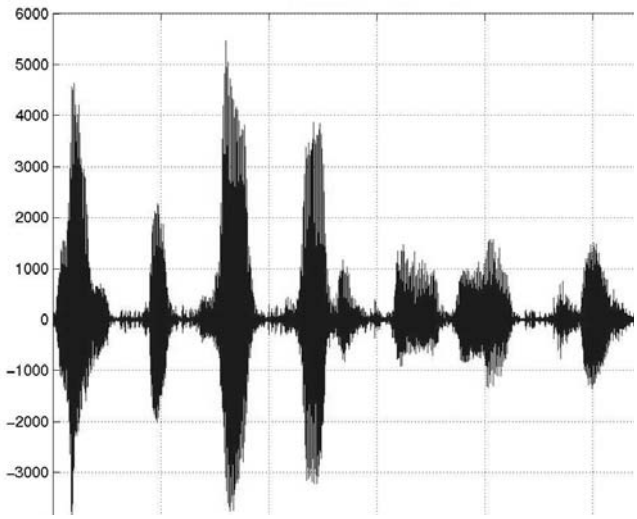

Figure 1.6: A typical speech signal

Voice biometrics has the most potential for growth, because it requires no new hardware, as most PCs already contain a microphone.However, poor quality microphones and ambient noise can severely affect verification (figure 1.6 shows a typical voice signal). In addition, the enrollment procedure has often been more complicated than with other biometrics, leading to the perception that voice verification is not user friendly. Therefore, current voice authentication systems still need improvement. One day, voice may become an additive technology to finger-scan technology. Because many people see finger scanning as a higher authentication form, voice biometrics will most likely be relegated to replacing or enhancing PINs, passwords, or account names.

Different biometric technologies may be appropriate for different applications,

depending on perceived user profiles, the need to interface with other systems or databases, environmental conditions, and a host of other application-specific parameters. Thus, there are a number of criteria to be argued, as shown in table 1.1, before deciding for or against a specific biometric technology.

Properties	Finger-print	Hand	Retina	Iris	Face	Signature	Voice
Ease of Use	High	V.High	Low	Medium	Medium	High	High
Accuracy	High	High	V.High	V.High	High	High	High
Cost	N.A	N.A	N.A	N.A	N.A	N.A	N.A
User Acceptance	Medium	Medium	Medium	Medium	Medium	V.High	High
Required Security Level	High	Medium	High	V.High	Medium	Medium	Medium
Long-term Stability	High	Medium	High	High	Medium	Medium	Medium

Table 1.1: Comparison of different biometrics [1]

Choice of a particular biometric is made depending upon the application of interest. For instance in a highly secured environment the issues of cost and ease of use will not have much weight, where as for a system which has to be implemented in a public place and is likely to be used more rapidly the user friendly environment would be a major requirement.

1.4 The Need for Multimodal Biometric Systems

The *monomodal* or *unimodal* biometric systems rely on the evidence of a single source of information for authentication (e.g., single fingerprint or face). These

systems have to deal with a variety of problems including,

1. **Noise in sensed data**: A fingerprint image with a scar, or a voice sample altered by cold are examples of noisy data. Noisy data could also result from defective or improperly maintained sensors (e.g., accumulation of dirt on a fingerprint sensor) or unfavorable ambient conditions (e.g., poor illumination of a users face in a face recognition system).

2. **Intra-class variations**: These variations are typically caused by a user who is incorrectly interacting with the sensor (e.g., incorrect facial pose), or when the characteristics of a sensor are modified during authentication (e.g., optical versus solid-state fingerprint sensors).

3. **Inter-class similarities**: In a biometric system comprising of a large number of users, there may be inter-class similarities (overlap) in the feature space of multiple users. Golfarelli et al. [5] state that the number of distinguishable patterns in two of the most commonly used representations of hand geometry and face are only of the order of 105 and 103, respectively.

4. **Non-universality**: The biometric system may not be able to acquire meaningful biometric data from a subset of users. A fingerprint biometric system, for example, may extract incorrect minutiae features from the fingerprints of certain individuals, due to the poor quality of the ridges.

5. **Spoof attacks**: This type of attack refers to the problem of forging the identity of a user. These attacks are likely when behavioral traits such as signature or voice are used. However, physical traits such as fingerprints are also susceptible to spoof attacks.

Some of the limitations imposed by unimodal biometric systems can be overcome by including multiple sources of information for establishing identity [6]. Such systems, known as *multimodal* biometric systems, are found more reliable due to the presence of multiple, (fairly) independent pieces of evidence [7]. These systems are able to meet the stringent performance requirements imposed by various applications. In this thesis, we show that combining speech and face does indeed improve recognition performance. Obviously, since independent systems are combined, we need to formulate a strategy for such combination. In this work, we propose to use the theory of evidence for such purpose.

1.5 Major Contributions of the Thesis

In this thesis, we have proposed a novel approach for multimodal biometric identification using the Dempster-Shafer theory of evidence (DST). In particular the major contributions of this thesis are summarized as follows:

1. Development of a new classifier combination algorithm called the **Nearest Neighbor based Evidence Fusion (NNEF)**, based on the DST for com-

bining "homogeneous" classifiers (Chapter 4).

2. Development of a new algorithm called the **Recognition Rate based Evidence Fusion (RREF)** algorithm based on the DST for combining heterogeneous multimodal classifiers (Chapter 5).

3. Development of a new algorithm called the **Variance based Evidence Fusion (VEF)** algorithm based on the DST for combining heterogeneous multimodal classifiers (Chapter 5).

4. Besides these major contributions we have also proposed: 1. a novel model based approach to the problem of face detection, and 2. a new neural network based algorithm for face recognition.

1.6 Organization of the Thesis

The thesis is organized as follows:

In **Chapter 2**, we propose a novel, model based face detection algorithm followed by a neural network based face recognition technique. Further more we discuss the issue of combining classifiers for pattern classification. We have discussed different categories of amalgamation of classifiers namely abstract level fusion, rank level fusion and measurement level fusion.

In **Chapter 3**, we give an introduction to the Dempster-Shafer theory of evidence

(DST). The conceptual difference between the Bayesian theory and the DST is discussed. The basic setup for the DST theory is formulated and the present methods for belief estimation are reviewed.

In **Chapter 4**, we propose the **Nearest Neighbor based Evidence Fusion algorithm (NNEF)** for fusing homogeneous classifiers. A DST based speaker recognition system is developed based on the NNEF algorithm.

In **Chapter 5** , we propose two different algorithms called the **Recognition Rate based Evidence Fusion (RREF)** algorithm and the **Variance based Evidence Fusion (VEF)** algorithm for combining heterogeneous classifiers. The algorithms are implemented for the case of multimodal biometrics (fusion of face and speech) and shown to outperform individual classifiers.

We conclude the thesis in **Chapter 6** with some concluding remarks, and propose some future research directions.

Chapter 2

Fundamental Concepts

2.1 Introduction

The ultimate goal of designing pattern recognition systems is to achieve the best possible classification performance for the task at hand. This objective traditionally led to the development of different classification schemes for any pattern recognition problem to be solved. The results of an experimental assessment of different designs can be used as the basis for choosing one specific classifier among many. It had been observed in such design studies, that although one of the designs would yield the best performance, the sets of patterns misclassified by the different classifiers would not necessarily overlap. This suggested that different classifier designs potentially offered complementary information about the patterns to be classified. Such complementary information could be harnessed to improve the performance of the selected

classifier. These observations motivated the relatively recent interest in combining classifiers. However before going into the details of the problem of fusing classifiers we choose a typical face recognition system and enhanced its performance using our own proposed approach. Before presenting the details of the proposed algorithm, it is worth noting that "face detection" is the first step towards the problem of face recognition. In many cases we first have to detect a human face in an image, then move to the recognition stage. With this understanding, we first address here the problem of human face detection as a background review and propose a model based technique for face detection. The work presented below has already been accepted for publication in [8].

2.2 A Typical Biometric System: Face Detection and Recognition

2.2.1 Frontal Human Face Detection in Complex Color Images

In this section we present a model based technique for extracting human faces, from complex, still color images. Most of the color images are represented in the RGB color space. RGB is not only a 3-dimensional space but also represents the brightness or luminance which is not a reliable criteria for skin separation due to the changed

ambient lightning [9]. To avoid the luminance and to reduce the color space, the RGB image is first transformed into the chromatic or pure color space [10, 11].

It has been observed that the skin colors of different people share almost the similar points in the color space. The difference in the apparent skin colors of different persons is mostly due to the intensity or luminance [12]. This fact is depicted in figure 2.1, which shows the clustering of skin color distribution for various skin colors.

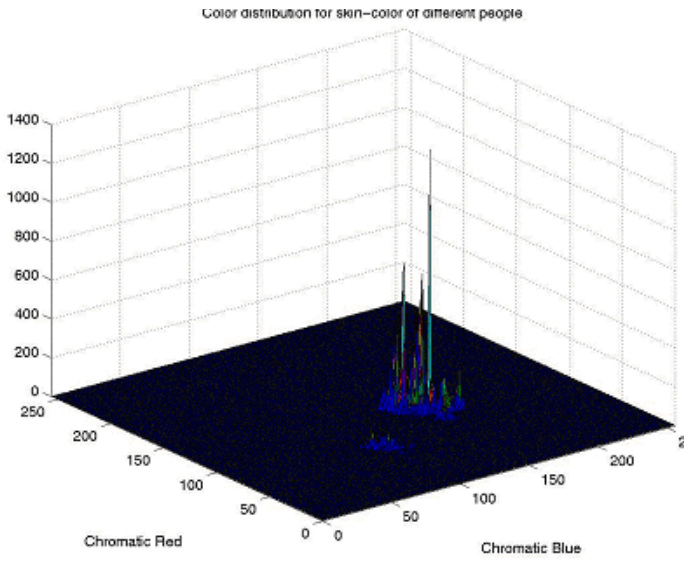

Figure 2.1: The color distribution for skin color of different people

Given the clustering of skin pixels around certain center point, we can reliably model the skin color using the Gaussian distribution [12] as shown in figure 2.2. The parameters used for such a Gaussian Model are as follows:

$$\mathbf{x} = [r \quad b]^T \tag{2.1}$$

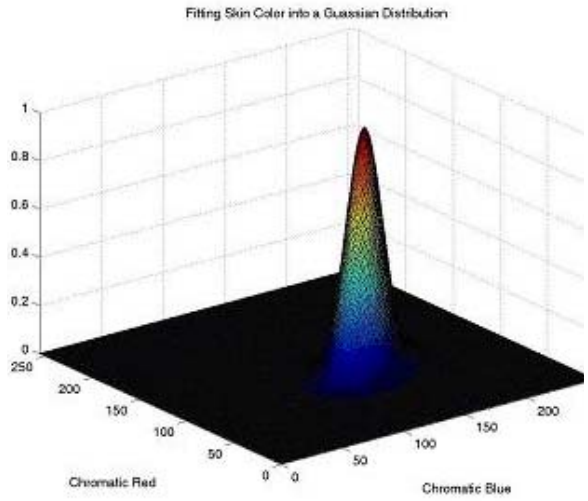

Figure 2.2: Gaussian model for skin

$$E(\mathbf{x}) = \mathbf{m} \tag{2.2}$$

$$\mathbf{C} = E[(\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^T] \tag{2.3}$$

where r and b are the components of the pure color space, $\mathbf{m}$ is the mean vector, and $\mathbf{C}$ is the covariance matrix of the Gaussian model. With this Gaussian distribution, the skin likelihood image can be obtained using the expression (up to a constant):

$$\mathbf{P}(r, b) = exp\Big[- 0.5(\mathbf{x} - \mathbf{m})^T \mathbf{C}^{-1} (\mathbf{x} - \mathbf{m}) \Big] \tag{2.4}$$

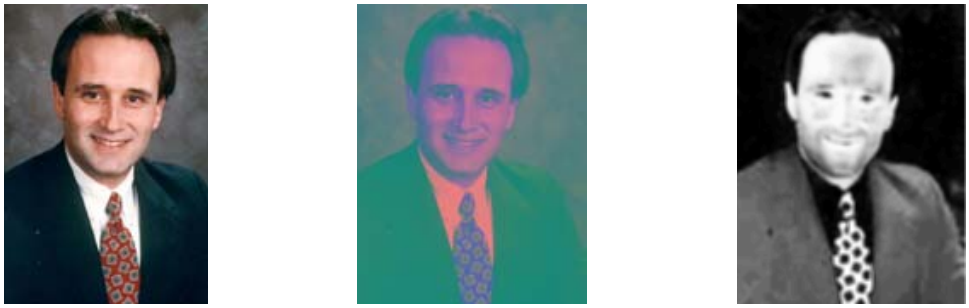

Figure 2.3: The original RGB image, the image transformed to chromatic color space and the skin likelihood image

Thus, we are able to express the chromatic or pure color representation into a gray scale image with brighter areas of the image showing the likelihood of the skin region, i.e the brighter a region, is the more likely it is to be a skin region. Figure 2.3 shows the transformation of an original image into the chromatic color space, then into the skin likelihood image.

After obtaining the skin likelihood image, we transform it into a binary image using an *adaptive thresholding* approach. It is obvious that if we decrease the threshold we will end up with an increased skin region. Thus, we propose here to decrease the threshold in steps and to select that value which gives minimum increase in the skin region. Once such threshold value is selected, all pixels having value above the threshold are designated as 1s and those having value less then the threshold are designated as non-skin (or 0s).

Since we are able to discriminate between the skin and non-skin regions, we must now check each of the skin regions for some characteristics so that we can decide which of the obtained region(s) correspond to a true human frontal face. In our

Figure 2.4: The resulting binary image

approach we explored the following characteristics:

- *Number of holes in a skin region:* It has been observed that a frontal human face will always have at least one hole (pixel value=0) corresponding to eyes etc. Thus to improve our decision process we have developed a criterion to reject all skin regions having no holes from being a human face candidate.

- *Height to width ratio:* Actually, human faces are vertically oriented [12] and ideally the height to width ratio is around 1.2. Thus, we can use this observation to classify that the regions having height to width ratio below 0.8 do not correspond to a human face. Similarly, we can put a higher upper limit on the ratio. However there are cases in which we have images with uncovered skin area below the face i.e neck etc, and to account for this, we put a higher upper limit of 1.6. Thus we would discard all those regions in our search of human face which have the region ratio less than 0.8 or above 1.6.

The most important development of the method is that it uses a template face

Figure 2.5: The average face

to match to the skin regions to make a final decision. The template face will be adapted using the geometric characteristics obtained for each region. It will first be resized using the height and width of the region. The resized template face is then oriented so that the template face has the same inclination as that of the region. The center of the inclined template face is then calculated and matched to the already calculated center of the region.

The cross-correlation between the adjusted template face and the skin region under consideration is then calculated, where the cross-correlation between any two functions $f(x, y)$ and $h(x, y)$ is given by:

$$\mathbf{E}[f(x,y); h(x,y)] = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f(m,n)h(x+m, y+n) \tag{2.5}$$

Empirically we have determined that a correlation value of 0.6 is good enough to decide that a given region corresponds to a human frontal face.

Figure 2.6 presents some of the results obtained using the proposed method.

The results show that the proposed method is robust for frontal human face

(a) The Skin Likelihood Image

(b) The Binary Image

(c) The Template Matching

(d) The Detected Face

Figure 2.6: The results of face detection for single face image

(a) The skin likelihood image

(b) The binary image

(c) The template matching

(d) The detected faces

Figure 2.7: The results of face detection for two face images

detection in complex color images. We have implemented the method for detection of multiple human faces in complex background and have found it extremely efficient. The developed technique is robust and efficient in the sense that it does not use complex neural networks, fuzzy integrals etc [13, 14, 15]

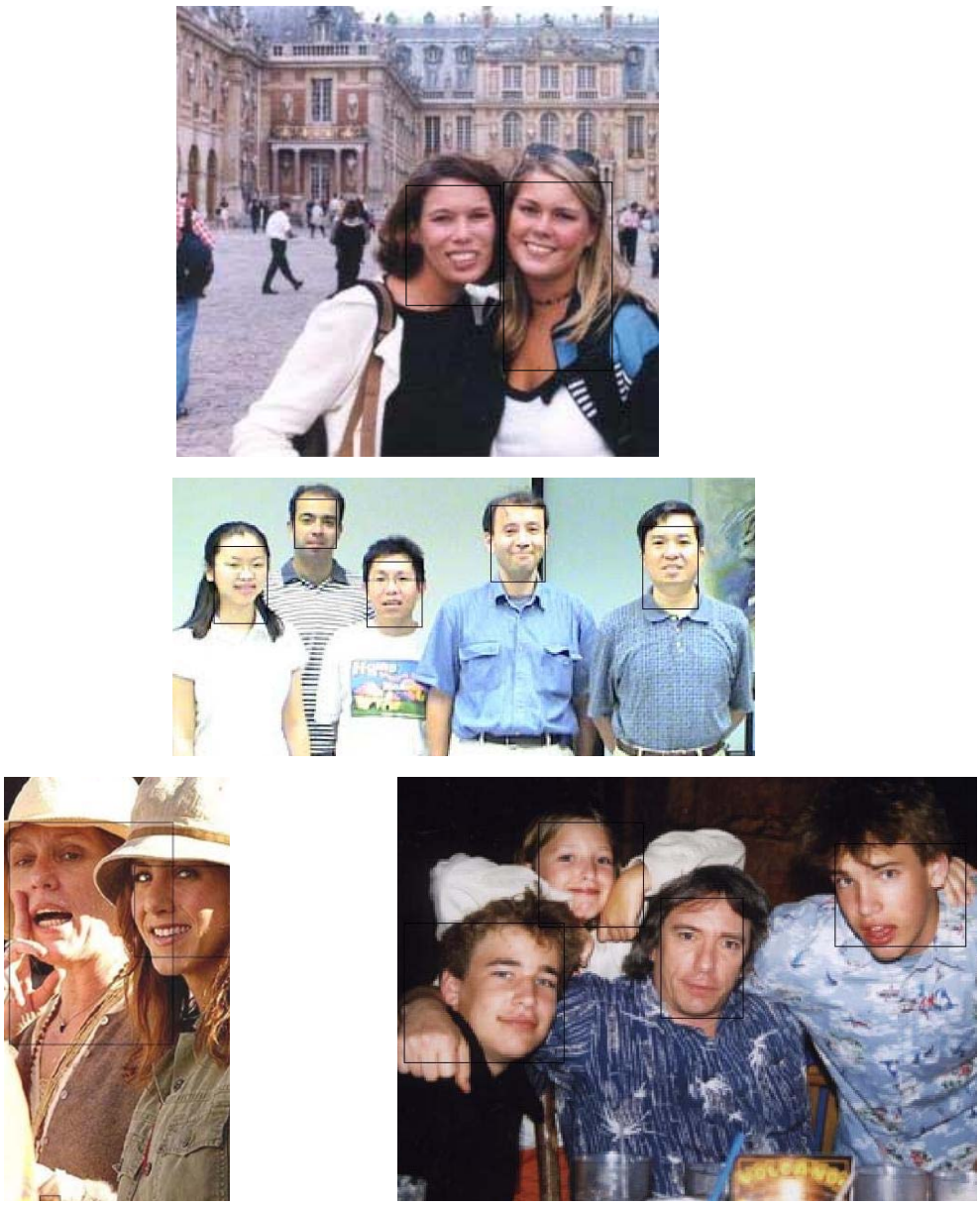

Figure 2.8: Some experimental results for the problem of face detection

2.2.2 Face Recognition

Face recognition has been a successful field of research mostly during the past two decades. The growth of research works in the field is mainly due to three factors:

1. The growing amount of face recognition applications reflected in the increasing number of face recognition companies.

2. The knowledge that face recognition models provide to the cognitive science field

3. The fact that face recognition has become a paradigm or benchmark of recognition methodologies.

In fact, face recognition, as a paradigm of a recognition system, has become a benchmark to the solutions of some of the main computer vision problems (invariance to view point; illumination change; occlusion; deformation due to changes of expression, age, make-up and hair style), as well to some of the main problems in statistical pattern recognition (feature selection; generalization; discriminability, etc.). This is evident when the continuous publication of reviews and surveys is considered, from the earliest of Samal and Iyengar (1992), to the latest of Jain (2003), passing through the works of Valentin et al. (1994), Chellappa et al. (1995) and Fromherz (1998) [16, 17, 18, 19]. Most of the statistical approaches to face recognition and detection, are based on Gaussian or mixture of Gaussian models [20, 21]. These methods are

mainly concerned with an estimation of the face manifold. However there are a few examples in which the neural network has been implemented for face recognition [22, 23]. These implementations have not been up to mark since they use small databases (less number of classes) and give low recognition rate. For instance in [22], 10 classes of AT&T database have been used giving a recognition rate of 75%. Similarly, in [23], results are shown only for 10 classes with a huge amount of training data. In this thesis, we have presented a new algorithm for face recognition using PCA (Principal Component Analysis) for feature extraction stage and neural networks for classification stage. We have shown that our system outperforms all conventional neural network based algorithms in both aspects of high recognition rate and lesser amount of required training data.

We start our discussion with a mathematical explanation of PCA or Principal Component Analysis which is used at the feature extraction stage.

2.2.3 Mathematical Analysis of PCA

The PCA or Principal Component Analysis is a technique to transform large dimensional data into a much reduced subspace called the "eigenspace". The main advantage of PCA is that it depicts all patterns in the data very efficiently. The variations in the data are modeled using the covariance matrix, we then perform eigenvalue eigenvector decomposition of such covariance matrix. These eigenvectors are the basis vectors and are used to transform the data into the eigenspace. The

mathematical implementation of the technique on images is as follows:

2-D facial images can be represented as 1-D vector by concatenating each row (or column) into a long thin vector. Lets suppose we have M vectors of size N (= rows of image x columns of image) representing a set of sampled images

$$\mathbf{x_i} = [p_1.....p_N]^T, i = 1, 2, ...M \tag{2.6}$$

where p_j represents a pixel value. The images are mean centered by subtracting the mean image from each of the image vectors. Let $\mathbf{m}$ represents the mean image:

$$\mathbf{m} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{x_i} \tag{2.7}$$

where M is the number of observations in the training data. Let $\mathbf{w_i}$s be defined as the mean centered images

$$\mathbf{w_i} = \mathbf{x_i} - \mathbf{m}, \quad i = 1, 2, ...M \tag{2.8}$$

Our goal is to find a set of vectors $\mathbf{e_i}$s which result in the largest possible variance of the vectors $\mathbf{w_i}$s. We wish to find a set of M orthonormal vectors $\mathbf{e_i}$s for which the quantity

$$\lambda_i = \frac{1}{M}\sum_{n=1}^{M}(\mathbf{e_i}^T\mathbf{w_n})^2 \tag{2.9}$$

is maximized with the orthonormality constraint

$$\mathbf{e_i}^T\mathbf{e_k} = \delta_{ik} \tag{2.10}$$

It has been shown that the $\mathbf{e_i}$s and λ_is are given by the eigenvectors and eigenvalues of the covariance matrix of the data, since the covariance matrix is not available, we

use an estimate of the covariance matrix:

$$\mathbf{C} = \mathbf{W}\mathbf{W}^{\mathbf{T}} \tag{2.11}$$

where $\mathbf{W}$ is a matrix composed of the column vectors $\mathbf{w_i}$ placed side by side. The size of $\mathbf{W}$ is NxN which could be enormous. For example, images of size 64 by 64 create the covariance matrix of size 4096 by 4096. It is not practical to find the eigenvectors of $\mathbf{C}$ directly. A common theorem in linear algebra states that the vectors $\mathbf{e_i}$s and scalars λ_is can be obtained by solving for the eigenvectors and eigenvalues of the MxM matrix $\mathbf{W}^T\mathbf{W}$. Let $\mathbf{d_i}$ and μ_i be the eigenvectors and eigenvalues of $\mathbf{W}^T\mathbf{W}$, respectively.

$$\mathbf{W}^T\mathbf{W}\mathbf{d_i} = \mu_i\mathbf{d_i}, i = 1, 2, ...M \tag{2.12}$$

or,

$$\mathbf{W}\mathbf{W}^T(\mathbf{W}\mathbf{d_i}) = \mu_i(\mathbf{W}\mathbf{d_i}) \tag{2.13}$$

which means that the first $M-1$ eigenvectors $\mathbf{e_i}$s and eigenvalues λ_is of $\mathbf{W}\mathbf{W}^T$ are given by $\mathbf{W}\mathbf{d_i}$ and μ_i, respectively. $\mathbf{W}\mathbf{d_i}$ needs to be normalized in order to be equal to $\mathbf{e_i}$. Since we only sum up a finite number of image vectors, M, the rank of the covariance matrix cannot exceed $M-1$ (the -1 comes from the subtraction operation of the mean vector $\mathbf{m}$).

The eigenvectors corresponding to the nonzero eigenvalues of the covariance matrix produce an orthonormal basis for the subspace within which most image data can

be represented with a small amount of error. The eigenvectors are sorted in order of decreasing eigenvalues, the eigenvector associated with the largest eigenvalue is one that reflects the greatest variance in the image. The eigenvalues decrease in an exponential fashion, with roughly 90% of the total variance contained in the first 5% to 10% of the first dimensions [20].

A facial image can be projected onto an M' ($M' \ll M$) dimension space by using

$$\mathbf{\Omega} = [v_1 v_2 ..... v_{M'}]^T \tag{2.14}$$

where $v_i = \mathbf{e_i}^T \mathbf{w_i}$ and v_i is the i^{th} coordinate of the facial image in the new feature space, which came to be the principal component. The vectors $\mathbf{e_i}$s are also images, so called, eigenimages, or eigenfaces in our case, which were first named in [20]. They can be viewed as images and indeed look like faces. PCA computes the basis of a space which is represented by its training vectors. When a particular face is projected onto the face space, its vector into the face space describes the importance of each of those eigenfaces in the overall face. The faces have representation in the face space by their eigenface coefficients (or weights). We can handle a large input vector, facial image, only by taking its small weight vector in the face space. This means that we can reconstruct the original face with minor errors, since the dimensionality of the image space is much larger than that of face space.

2.2.4 Neural Networks in Face Recognition

Artificial Neural Networks (ANNs) are computational systems with architecture and operation inspired from our knowledge about biological neural cells (neurons) in the brain. ANNs can be described either as mathematical and computational models for non-linear function approximation, data classification, clustering and non-parametric regression, or as simulations of the behavior of collections of model biological neurons. These are not simulations of real neurons in the sense that they do not model the biology, chemistry, or physics of a real neuron. They do, however, model several aspects of the information combining and pattern recognition behavior of real neurons in a simple yet meaningful way. Neural modeling has shown incredible capability for emulation, analysis, prediction, and association. ANNs have also been used in a variety of powerful ways: to learn and reproduce rules or operations from given examples; to analyze and generalize from sample facts and make predictions from these; to memorize characteristics and features of given data, and to match or make associations from new data to the old data.

ANNs can be used to solve difficult problems in a way that resembles human intelligence. What is unique about neural networks is their ability to learn by example. Traditional artificial intelligence (AI) solutions, on the other hand, rely on symbolic processing of the data, an approach which requires a priori human knowledge about the problem. Neural networks techniques have also an advantage

over statistical methods of data classification as they are distribution-free and require no a priori knowledge about the statistical distributions of the classes in the data sources in order to classify these. Unlike the statistical approaches, ANNs are able to solve problems without any a priori assumptions. As long as enough data is available, a neural network can learn and extract any regularities and form a solution.

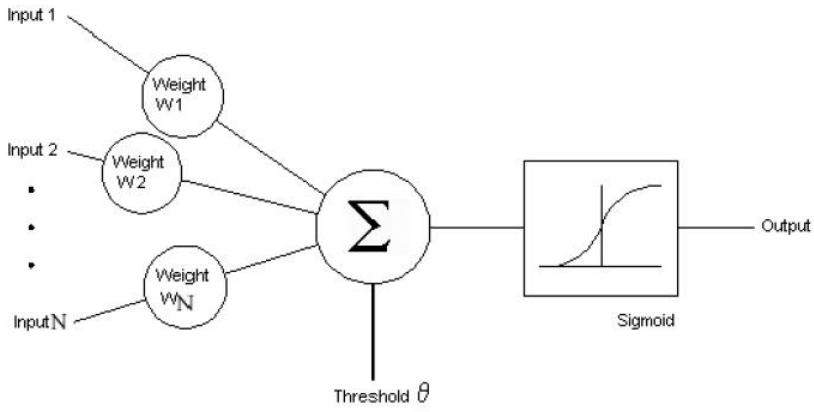

Figure 2.9: The basic structure of a neuron [2]

As ANNs are models of biological neural structures, the starting point for any kind of neural network analysis is a model neuron whose behavior follows closely our understanding of how real neurons work. This model neuron is shown in figure 2.9. The neuron has N input lines and a single output. Each input signal is weighted, that is, it is multiplied with the weight value of the corresponding input line (by analogy to the synaptic strength of the connections of real neurons). The neuron combines these weighted inputs by forming their sum and, with reference to a threshold value and activation function, it determines the output. In mathematical terms, we may

describe the neuron by writing the following pair of equations:

$$u = \sum_{i=1}^{N} w_i x_i \tag{2.15}$$

$$y = f(u - \theta) \tag{2.16}$$

where x_is are the inputs to the neuron, w_is are the weights of the neuron, u is the weighted sum, $f(.)$ is the nonlinearity function and y is the final output.

Feed-forward networks form the most important class among the different classes of neural networks. Typically, these consist of a set of sensory units (source nodes) that constitute the input layer, one or more hidden layers of computation nodes and an output layer of computation nodes. The input signal propagates through the network in a forward direction, on a layer-by-layer basis. These neural networks are commonly referred to as *multilayer perceptrons*(MLPs) (figure 2.10). MLPs have been applied successfully to solve some difficult and diverse problems by training these in a supervised manner with the highly popular algorithm known as the *error back-propagation algorithm.*

2.2.5 The Proposed Algorithm and Obtained Results

In our proposed system, after the face is detected, the PCA features are extracted, followed by classification using an MLP based ANN. Extensive experiments were carried out using the proposed algorithm. Eigen space dimension was chosen to be

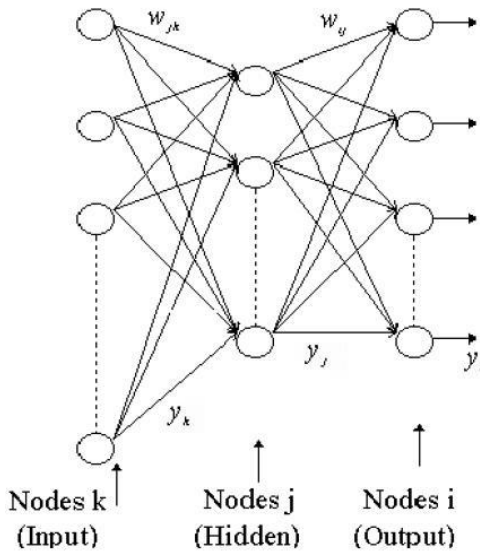

Figure 2.10: Topology of a single hidden layer MLP [2]

10, since it has been observed that increased dimension does not contribute to the accuracy [20]. A single hidden layer MLP based architecture of neural network is implemented with back propagation method used for weight updating.

Table 2.1 shows the results using the AT&T database. The pose variations for one class has also been shown. We have taken 10 classes i.e there were 10 persons, with a training set of 7 images and testing set of 3 images per person. The results are displayed in terms of cost function which is a measure of performance of learning, achieving an overall recognition rate of 93.3%. Our algorithm outperformed the recognition rate of 75% on AT&T database for same number of classes as achieved in [22].

CLASS	Recognized as									
	1	2	3	4	5	6	7	8	9	10
1	3	0	0	0	0	0	0	0	0	0
2	0	3	0	0	0	0	0	0	0	0
3	0	0	3	0	0	0	0	0	0	0
4	0	0	0	3	0	0	0	0	0	0
5	0	0	0	1	2	0	0	0	0	0
6	0	0	0	0	0	3	0	0	0	0
7	0	0	0	0	0	0	3	0	0	0
8	0	0	0	0	0	0	0	3	0	0
9	0	0	0	0	0	0	0	0	3	0
10	0	0	0	0	0	0	0	1	0	2

Table 2.1: Classifications results with the AT&T database

CLASS	TRUE CLASSIFICATIONS	FALSE CLASSIFICATIONS
1	3	0
2	3	0
3	2	1
4	3	0
5	3	0
6	3	0
7	3	0
8	1	2
9	2	1
10	3	0
11	3	0
12	3	0
13	3	0
14	3	0
15	3	0

Table 2.2: Classification results with the YALE database

The algorithm was also tested using the YALE database (15 classes with 11 images per class) with 8 training images and 3 testing images per class. The results of classification for each class are shown in Table 2 with an overall recognition rate of 91.11%. An example of misclassification result is shown in figure 2.12. The main contribution of the approach has been the improvement achieved in recognition rate. Note that we have performed experiments without any thresholding, thus the issue of false acceptance rate is not addressed here.

2.3 Combining Multiple Classifiers

The idea of combining multiple classifiers is not to rely on a single decision making scheme. Instead, all the designs, or their subsets, are used for decision making by combining their individual "opinions" to derive a consensus decision. Various classifier combination schemes have been devised and shown to lead to better classification results then those obtained using a single best classifier. However, there is presently inadequate understanding why some combination schemes are better than others and in what circumstances. It is mentioned in [4] that there are two main reasons for combining classifiers: efficiency and accuracy.

The existing classifier combination methods can be divided into two groups: the multi-stage methods and the ensemble methods. The multi-stage methods decompose the classification problem into a set of subproblems that can be solved by

Figure 2.11: A subject of AT&T database with various poses.

(a) The unknown image (b) The recognized image

Figure 2.12: An example of misclassification

"local" classifiers. Therefore, the focus is more on the decomposition rather than on the classifiers. On the other hand, the main point for ensemble methods is that a group (ensemble) of "global" classifiers potentially gives better generalization than the individual classifiers (ensemble members).

The fusion of classifiers could be achieved at three levels i.e data level fusion, feature level fusion and decision level fusion. In this thesis, we have addressed the decision level fusion problem.

2.4 The Multi-Stage Classifiers

The multi-stage classifier architecture is based on the divide-and conquer principle, in which a large, hard to solve problem is broken up into many smaller, easier to solve problems. This principle yields good performance and allows fast training. Many types of classifiers have been used in multi-stage methods. Simple classifiers were used in [24], generalized linear models were used in [25], and ANNs were used in [26]. An example of multi-stage classifiers was given by Cao et al. [27], where the problem of handwritten numerals recognition was considered. Rather than classifying the patterns into 10 classes in one step, a subclass method of 2 classes was used. Thus, the 10-category numeral recognition problem was divided into 45 2-category classification problems, where each sub-classifier was for two numerals. In the first stage, a clustering neural network decided which classifier to be used for an unknown

input pattern. Zhou and Pavlidis [28] proposed a hierarchical character recognition scheme. An object was first classified based on a preliminary shape description and then justified the shape on questionable parts according to acquired class knowledge and additional information. Polygonal features and contour features were used as preliminary and secondary sources of shape information. Other examples of multi-stage methods can be found in [29, 30, 31, 32, 33, 34, 35]. The multi-stage classifiers are often used to solve specific problems [36].

2.5 Combination of a Group of Global Classifiers

Combining classification results of a group of global classifiers has recently received considerable attention as a new direction for the development of highly reliable pattern classification systems. This is due to the following:

1. In many pattern classification problems, there are a number of classification algorithms available. These algorithms are based on different theories and methodologies. For a specific problem, these classifiers usually attain different degrees of success, but the perfection of one technique cannot be claimed. We need to investigate ways of integrating the results of these different classifiers in order to achieve better results.

2. Many types of features have been proposed to represent patterns. These features could be in the form of binary values, discrete labels, continuous vari-

ables, etc. No single classifier can handle all types of features, and the only way to process them is by using many classifiers. Hence, we need to integrate the results obtained by classifying the different types of features to obtain better results than any specific set of features can provide.

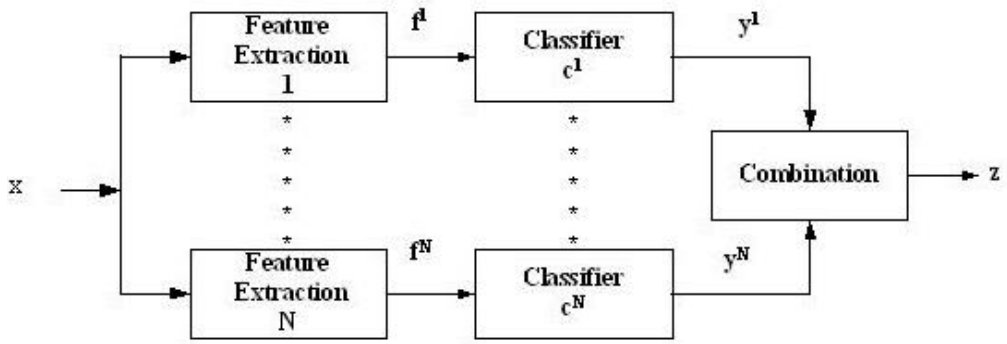

Figure 2.13: Multiple classifier system

In other words, the use of multiple classifier systems is motivated by the existence of many alternative solutions to a pattern classification problem, and the observation that these solutions often complement one another in accuracy. A block diagram of a multiple classifier system is shown in figure 2.13.

An interesting question is whether it is possible to integrate these alternative solutions, in such a way that the integration excels the individual solutions in performance. In several preliminary studies [37, 38, 39, 40, 41], there are hints that robust solutions to certain recognition problems may involve a number of independent methods. These studies suggest the idea of a multiple classifier system. The

main issue that needs to be tackled in designing a multiple classifier system is the conflict between classifiers, which arises when two or more classifiers make different decisions. As explained in [42], the problem of combining multiple classifiers consists of two parts. The first part, closely dependent on specific applications, includes the problems of "How many and what type of classifiers should be used for a specific application?, and for each classifier what type of features should we use?", as well as other problems that are related to the construction of those individual and complementary classifiers. For the particular case of handwritten numerals, good description of these issues can be found in [43, 44], while Alkoot and Kittler [45] studied the effect of adding classifiers to the multiple classifier architecture and adding new features to each of the classifiers in the architecture. The second part, which is common to various applications, tackle the problems related to the question "How to combine the results from different existing classifiers so that a better result can be obtained?". In our work, we will be concentrating on problems related to the second part. It is first necessary to understand the type of information provided by each of the classifiers before trying to combine their results. The output information from various classification algorithms can be categorized into three levels [42, 46]:

1. The abstract level: a classifier only outputs a unique label, as in the case of syntactic classifiers.

2. The rank level: a classifier ranks all class labels or a subset of the class labels

in a queue with the label at the top being the first choice.

3. The measurement level: a classifier attributes to each class a measurement value that reflects the degree of confidence that a specific input belongs to a given class. This degree could be a probability, as in the Bayesian classifier, or any other scoring measure.

It is obvious that class ranking can be obtained from class measurements, as well as class labels can be obtained from the top choice among class ranking and class measurements. In other words, the measurement level contains the highest amount of information while the abstract level contains the lowest.

The basic formulation of the problem is as follows: Consider that we have N different classifiers, $c^n, n = 1, ..., N$. Then the input pattern $\mathbf{x}$ is assigned to one of the K possible classes $\{\omega_1, ...\omega_k, ...\omega_K\}$ by each classifier. The input feature vector to classifier c^n is $\mathbf{f^n}$, and the type of classification vector $\mathbf{y}^n = [y^n(1), ..., y^n(K)]^T$ produced by c^n could be either abstract, ranking, or measurement. Below are descriptions of the different combination methods developed for the three decision levels.

2.6 Combination of Abstract level Decisions

A number of methods have been used to combine classification results at the abstract level. Below are the most well-known techniques:

2.6.1 Majority Voting

The majority voting can be considered as the default combination scheme for this type of information. It has received great attention as it is seen as the simplest combination scheme and lends itself easily to theoretical analysis of behavior and performance [47, 41, 48, 49, 4].

As indicated by its name, the method is based on the majority voting principle. If classifier c^n assigns a given pattern to class label ω_k, then we say that a vote is given to ω_k. After counting the votes given to each class label by all classifiers, the class label that receives a number of votes higher than others (or higher than a prefixed threshold) is taken as the final output.

Hansen and Salamon [41] showed that if independent neural networks are combined, provided that each network can get the right answer more than half the times, then the more networks used, the less is the likelihood of an error. A weighted voting scheme was also proposed in [50]. The vote of each classifier was weighted according to its performance, i.e., the weight of a good classifier would be higher than that of a less powerful classifier. Thus, this approach takes into account the reliability of the different classifiers. Lam and Suen [51] proposed a method to obtain the weights of each classifier through the optimization of an objective function for the combined decision. Bayesian formulation and genetic algorithms were also studied. Xu et al. [42] described the different variants of the voting principle and presented a general

formula to the voting function.

2.6.2 Bagging and Boosting

Two decision expert learning approaches, boosting [52, 53] and bagging [54], have received extensive attention recently [55, 56]. The Bagging algorithm (Bootstrap aggregating) votes classifiers generated by different bootstrap samples. A Bootstrap sample is generated by uniformly sampling m instances from the training set with replacement. T bootstrap samples $B_1, B_2, ..., B_T$ are generated and a classifier C_i is built from each bootstrap sample B_i. A final classifier $C*$ is built from $C_1, C_2, ..., C_T$ whose output is the class, predicted most often by its sub-classifiers, with ties broken arbitrarily.

Boosting was introduced by Schapire (1990) as a method for boosting the performance of a weak learning algorithm. Like Bagging the Boosting algorithm generates a set of classifiers and votes them. It changes the weight of the votes given to each classifier in an adaptive manner. Given an integer T specifying the number of trials, T weighted training sets $S_1, S_2, ..., S_T$ are generated in sequence and T classifiers $C_1, C_2, ..., C_T$ are built. A final classifier is formed using a weighted voting scheme: the weight of each classifier depends on its performance on the training set used to build it.

There are two major differences between bagging and boosting. First, boosting changes adaptively the distribution of the training set based on the performance of

previously created classifiers while bagging changes the distribution of the training set stochastically [57]. Second, boosting uses a function of the performance of a classifier as a weight for voting, while bagging uses equal weight voting [57]. With both techniques the error decreases when the size of the experts increases, but the marginal error reduction of an additional member tends to decrease [57].

2.6.3 Dempster-Shafer Theory of Evidence

Dempster-Shafer theory of evidence or the DST, proposed in the mid of 1970, has shown its power for modeling uncertainty. In order to take into account uncertainties in classification, the Dempster-Shafer theory of evidence was used in [42] to combine the classification results. The recognition, substitution, and rejection rates were used to measure the belief of each classifier. When tested experimentally, this method was found to be quite robust, and was shown to outperform majority voting. However, the way the belief was measured in [42] is not optimal, as it does not take into consideration the accuracy with respect to each class label, and hence does not resolve conflicts between classifiers in an optimal way.

2.6.4 Bayesian Formulation

The confusion matrix, which describes the errors of different classifiers, was used by Xu et al. [42] to estimate the conditional probabilities that input pattern $\mathbf{x}$ belongs to class label ω_k is true under the condition that classifier c^n chooses label ω_j (i.e.,

$y^n(j) = 1)$, that is:

$$P(x \in \omega_k / y^n(j) = 1) = \frac{e_{kj}^{(n)}}{\sum_{k=1}^{K} e_{kj}^{(n)}} \tag{2.17}$$

where $e_{kj}^{(n)}$ is the number of patterns actually belonging to ω_k that have been assigned to ω_j by the classifier c^n. On the basis of the above probabilities, the combination of independent classifiers can be carried out by multiplying the normalized conditional probabilities. The pattern is then assigned to the class for which the combined value is maximum. A detail of the Bayesian approaches could be found in [58].

2.6.5 Behavior-Knowledge Space

Huang and Suen [59] presented a combination scheme designed to avoid the implications of the independence assumption. To this end, prior knowledge on the behavior of the classifier ensemble is recorded in the behavior-knowledge space (BKS). It is an N-dimensional space where each dimension corresponds to the decision of one classifier. Each intersection in the BKS represents a possible N-tuple of decisions taken by the classifiers in the ensemble. The number of samples of each class with the same ensemble behavior is stored in each BKS intersection. When a new pattern needs to be classified, the answers of the ensemble are used to pick the proper entry in the BKS. The class that accumulated the highest number of samples in the picked BKS entry is then taken as the true class for the test pattern. Since the independent

assumption of Bayesian Formalism fails to hold for many real situations, and BKS is hard to implement, due to its computational and large storage requirements, Kang et al. [60, 61] proposed an intermediate approach based on the approximation of the product of Nth order dependency.

2.7 Combination of Rank Level Decisions

The rank level classifiers give an output vector with class labels arranged in a decreasing order of priority. Little work has been conducted in the combination of rank level decisions mainly because the information provided by most classifiers falls either into the abstract or measurement levels. The main contribution was made by Ho et al. [62], where two classes of techniques have been investigated: class set reduction and class set reordering. Class set reduction aims at reducing the number of classes in the output list without losing the true class. The criteria of success are: the size of the result set should be minimized, and the probability of the inclusion of the true class should be maximized. The method derives a threshold on the ranks according to the worst-case ranks of the true classes. On the other hand, class set reordering attempts to improve the rank of the correct class. The criterion of success is the position of the true class in the resultant ranking, as compared to its position in the ranking before combination. A method is considered successful if the probability of having the true class near the top of the combined ranking is

higher than that of each of the original rankings.

The use of such methods allows the combination of classifiers where the outputs are at the measurement level but they are expressed on different scales. Rankings can be in fact easily derived from measurement outputs. Of course some information would be lost since the confidence of the decisions cannot be estimated.

2.8 Combination of Measurement Level Decisions

Measurement level classifiers give output in form of confidence (score) in each class label. Because classifiers that produce measurement decisions are widely used, many researchers have concentrated on the combination of such classifiers. Below are the well-known methods developed to date:

2.8.1 Traditional Methods

A theoretical framework for combining classifiers using traditional methods was developed by Kittler et al. [4]. Classifier combination strategies based on the product and sum decision rules were investigated. The max,min,median and majority voting rules were derived from the above two rules.

Previous experimental comparison of various classifier combination schemes showed that the combination under the sum rule outperformed the other classifier combination scheme. Alkoot and Kittler [63] emulated the behavior of individual classifiers

by subjecting their nominal soft outputs to perturbation errors. The experiments showed that combined classifiers give better results than those by single classifier, especially the sum and median. However, the single classifier might be preferable over the product, minimum, and maximum, under the Gaussian noise assumption of the estimation error. The results also confirmed the theoretical prediction that in most scenarios the sum rule outperformed the product rule and other strategies derived from it [63].

2.8.2 The Dempster-Shafer Theory of Evidence

Because of its ability in representing uncertainty and lack of knowledge, the Dempster-Shafer theory of evidence has generated considerable interest in various fields, including classifier combination. Measurement outputs could provide more useful information, compared to abstract and rank outputs, that can help in estimating the evidence of classifiers more accurately. Mandler and Schurmann [40] attempted to estimate the a posteriori probability function for both intra and interclass distances, which were then transformed into evidence. But due to the approximations associated with estimating the statistical models of intra and interclass distances, the accurate estimation of evidence could not be claimed. Rogova [64] proposed to use a reference vector and a proximity measure to estimate the evidence. The results obtained were promising. However, the techniques to appropriately choose both the reference vector and the proximity measure were not well formulated.

2.8.3 Re-classifying the original classification results

Stacked generalization, a scheme for minimizing the generalization error rate of one or more classifiers, was proposed by Wolpert [65]. Stacked generalization works by deducing the biases of the classifier(s) with respect to a provided learning set.

Merz [66] used the strategies of stacking and correspondence analysis to model the relationship between the learning examples and their classification by a collection of learned classifiers. Correspondence analysis is a method for geometrically modeling the relationship between the rows and columns of a matrix whose entries are categorical [67]. The goal was to explore the relationship between the training examples and their classification by the learned models. In an empirical analysis, the method was showed to be insensitive to poor learned models and matched the performance of plurality voting as the errors of the learned models become less correlated.

Krogh and Vedelsby [68] defined the ambiguity as the variation of the output of ensemble members averaged over unlabeled data, which quantifies the disagreement among classifiers. They used the ambiguity with cross-validation to give an estimate of the ensemble generalization error.

A two-fold approach to the problem that contains data transformation and data classification was proposed by Huang et al. [69]. In data transformation, the classification results of each classifier are transformed into a form of likeness measurement.

The larger the likeness measurement is, the more probable the corresponding class belongs to that input. In data classification, a neural network was used to aggregate the transformed results to produce the final classification decisions.

2.9 Chapter Summary

In this chapter we have proposed a neural network based face recognition system which outperforms the previous benchmark algorithms. With the understanding that in a real time implementation, a localized face is not always available, we developed a robust model based technique for face detection in complex images. Further, we discussed the issue of combining classifiers. Classifier designs differ from each other in two aspects:

1. The method in which the features are extracted and

2. The algorithm used for classification.

Different classifier designs offer complementary information for a given test pattern. This observation provides enough rational to conduct research in area of combining the complementary classifiers so that a consensus decision is achieved. Studies have shown that a combined classifier system consistently outperforms the individual classification results.

We have discussed different methods of combining classifiers for e.g majority voting, Bayesian formulation, Dempster-Shafer theory of evidence etc. Among these

the Dempster-Shafer theory of evidence models the uncertainty in a better way by avoiding the over commitment in certain hypothesis. In the next chapter, we discuss the Dempster-Shafer theory of evidence and its implementation to the classifier combination problem with application to multimodal biometrics.

Chapter 3

Combining Classifiers Using the Dempster-Shafer Theory of Evidence

3.1 Introduction

In pattern recognition, the main objective is to achieve the highest possible classification accuracy. To attain this objective, researchers, throughout the past few decades, have developed numerous systems working with different features depending upon the application of interest. These features are extracted from the data and can be of different types like continuous variables, binary values, etc. As such, a certain classification algorithm used with a specific set of features may or may

not be appropriate with a different set of features. In addition, classification algorithms are different in their theories, and hence achieve different degrees of success for different applications. Even though, a specific feature set used with a specific classifier might achieve better results than those obtained using another feature set and/or classification scheme, one can not conclude that this set and this classification scheme achieve the best classification results [46]. It has been found that the different classifiers used for a special classification task usually complement each other with respect to the information extracted from the patterns to be classified [4]. As a result, combining the different classifiers, in an efficient way, is expected to achieve better classification results than any single classifier including the best one.

As explained in [42], the problem of combining multiple classifiers consists of two parts. The first part includes the problems of "How many and what type of classifiers should be used for a specific application?, and for each classifier what type of features should we use?", as well as other problems that are related to the construction of these individual and complementary classifiers. The second part covers the problems related to the question "How to combine the results from different existing classifiers so that a better result can be obtained?". Several combination methods based on different theories have been proposed in the literature however one of the important issues that needs to be considered when combining classifiers is the level of uncertainty associated with the performance of each of the classifiers. In the following section, we discuss how the Dempster-Shafer theory of evidence is

an appropriate approach when it comes to representing uncertainty.

3.2 Representation of Uncertainty

Let us first give a simple example to explain uncertainty. Let θ represent the following proposition: the *passionfruit* is delicious. Then according to the Bayesian theorem $P(\theta) + P(\overline{\theta}) = 1$, where $\overline{\theta}$ is negation of θ. Now suppose that Jim has not tasted the passionfruit before. Then, we cannot say that Jim believes the proposition if he has no idea what it means. Also, it is not fair to say that he disbelieves the proposition. This problem can be better represented by the Dempster-Shafer (D-S) theory of evidence, which is regarded as a more general approach to representing uncertainty than the Bayesian approach. The D-S theory would denote Jim's belief of the proposition, $m(\theta)$, and disbelief, $m(\overline{\theta})$, as both being zero. Certainty factors do not allow this.

Thus, the difference between the Bayesian statistical model and the D-S evidential theory is conceptual. In the statistical model, it is assumed that there is a Boolean phenomena which either does or does not exist. The result of this assumption leads to the implication that commitment of belief to a hypothesis leads to the commitment of the remaining belief to its negation. If there is little belief for the existence of a phenomena this would imply, under the Bayesian formulation, a large belief to its non-existence, which is what we call *over-commitment*. In D-S

theory, one considers the evidence in favor of a hypothesis. There is no causal relationship between a hypothesis and its negation, hence lack of belief does not imply disbelief. Rather, lack of belief in any particular hypothesis implies belief in the set of all hypotheses, which is referred to as the state of uncertainty. If we denote the uncertainty by Θ, then in the above example $m(\Theta) = 1$, which is calculated by the following formula: $m(\theta) + m(\overline{\theta}) + m(\Theta) = 1$. For this reason, we will only be concerned here with the D-S theory of evidence and its application to the problem of classifier combination.

We now introduce some basic concepts of the Dempster-Shafer theory of evidence.

3.3 The Dempster-Shafer Theory of Evidence

The D-S theory of evidence [70] is a powerful tool for representing uncertain knowledge. This theory has inspired many researchers to investigate different aspects related to uncertainty and lack of knowledge and their applications to real life problems [71]. Today, the D-S theory covers several different models including the transferable belief model (TBM) [70].

In order to explain the combination rule under the TBM model, we need to present the definitions of basic belief assignment and belief function. Let $\Theta = \{\theta_1, ...\theta_K\}$ be a finite set of possible hypotheses. This set is referred to as the frame

of discernment, and its powerset denoted by 2^{Θ}. The basic belief assignment of a subset of Θ and the belief function associated with it are defined as follows:

3.3.1 Basic Belief Assignment (BBA)

A basic belief assignment $m(.)$ is a function that assigns a value in [0,1] to every subset $\mathbf{A}$ of Θ and satisfies the following:

$$m(\phi) = 0, and \sum_{\mathbf{A} \subseteq \Theta} m(\mathbf{A}) = 1 \tag{3.1}$$

where ϕ is the empty set. It is worth noting that $m(\phi)$ can be non-zero when considering un-normalized combination rule as will be explained later. While in probability theory a measure of probability is assigned to atomic hypotheses θ_i, $m(\mathbf{A})$ is the measure of belief that supports $\mathbf{A}$, but does not support anything more specific, i.e., strict subsets of $\mathbf{A}$. For $\mathbf{A} \neq \theta_i$, $m(\mathbf{A})$ reflects some ignorance because it is a belief that we cannot subdivide $\mathbf{A}$ into finer subsets. $m(\mathbf{A})$ is a measure of support we are willing to assign to a composite hypothesis $\mathbf{A}$ at the expense of support $m(\theta_i)$ of atomic hypotheses θ_i. For a particular frame of discernment Θ, if we set $m(\theta_i \neq 0)$ for all θ_i and $m(\mathbf{A}) = 0$ for all A $\neq \theta_i$, then $m(\theta_i)$ becomes probability of θ_i with $\sum_i m(\theta_i) = 1$. A subset $\mathbf{A}$ for which $m(\mathbf{A}) > 0$ is called a focal element. The partial ignorance associated with $\mathbf{A}$ leads to the following inequality: $m(\mathbf{A}) + m(\overline{\mathbf{A}}) \leq 1$, where $\overline{\mathbf{A}}$ is the complement of $\mathbf{A}$. In other words, the D-S theory of evidence allows us to represent only our actual knowledge without

being forced to *overcommit* when we are ignorant.

3.3.2 Belief Function

The belief function, $bel(.)$, associated with the BBA $m(.)$ is a function that assigns a value in [0,1] to every nonempty subset $\mathbf{B}$ of Θ. It is called *degree of belief in* $\mathbf{B}$ and is defined by

$$bel(\mathbf{B}) = \sum_{\mathbf{A} \subseteq \mathbf{B}} m(\mathbf{A}) \tag{3.2}$$

where $\mathbf{A}$ is subset of $\mathbf{B}$. We can consider a basic belief assignment as a generalization of a probability density function whereas a belief function is a generalization of a probability distribution function.

3.3.3 Combination rule

Consider two BBAs $m_1(.)$ and $m_2(.)$ for belief functions $bel_1(.)$ and $bel_2(.)$ respectively. Let $\mathbf{A}_j$ and $\mathbf{B}_k$ be focal elements of $bel_1(.)$ and $bel_2(.)$ respectively. Then $m_1(.)$ and $m_2(.)$ can be combined to obtain the belief committed to $\mathbf{C} \subset \Theta$ according to the following combination or *orthogonal sum* formula [70],

$$m(\mathbf{C}) = m_1(\mathbf{C}) \oplus m_2(\mathbf{C}) = \frac{\sum_{j,k,\mathbf{A}_j \cap \mathbf{B}_k = \mathbf{C}} m_1(\mathbf{A}_j) m_2(\mathbf{B}_k)}{1 - \sum_{j,k,\mathbf{A}_j \cap \mathbf{B}_k = \phi} m_1(\mathbf{A}_j) m_2(\mathbf{B}_k)}, \ \mathbf{C} \neq \phi \tag{3.3}$$

The denominator is a normalizing factor, which intuitively measures how much

$m_1(.)$ and $m_2(.)$ are conflicting. Smets [72] proposed the un-normalized combination rule :

$$m_1(\mathbf{C}) \cap m_2(\mathbf{C}) = \sum_{j,k,\mathbf{A}_j \cap \mathbf{B}_k = \mathbf{C}} m_1(\mathbf{A}_j) m_2(\mathbf{B}_k) \, , \forall \mathbf{C} \subseteq \Theta \tag{3.4}$$

This rule implies that $m(\phi)$ could be positive, and in such case reflects some kind of contradiction in the belief state. In our passionfruit example, suppose that Kim has tasted the passionfruit and expressed his belief as follows: $m_2(\theta) = 0.8$, $m_2(\overline{\theta}) = 0.2$ and $m_2(\Theta) = 0$. The reason behind assigning 0 to $m_2(\Theta)$ is that Kim knows exactly what the proposition means. Combining the beliefs of Jim and Kim according to the combination rule would lead to: $m(\theta) = 0.8$, $m(\overline{\theta}) = 0.2$ and $m(\Theta) = 0$, which is Kim's belief. This makes sense, as there is no reason for the totally uncertain belief to have any effect on the combination outcome. Now assume that Lim has only tasted the passionfruit once while eating a fruit salad and that he did not like the taste of the fruit salad. Lim expressed his belief as follows: $m_3(\theta) = 0.1$, $m_3(\overline{\theta}) = 0.4$ and $m_3(\Theta) = 0.5$. Lim was a bit uncertain because he had no clear idea about the taste of the passionfruit itself. The outcome of combining Kim's and Lim's beliefs would be: $m(\theta) = 0.73$, $m(\overline{\theta}) = 0.27$ and $m(\Theta) = 0$. Note how the result is influenced by Kim's belief. This is due to the uncertainty of Lim, without which, both Kim and Lim would have the same influence on the combination outcome.

3.3.4 Combining Several Belief Functions

The combination rule can easily be extended to several belief functions by repeating the rule for new belief functions. Thus the pairwise orthogonal sum of n belief functions $bel_1, bel_2, ..., bel_n$, can be formed as

$$((bel_1 \oplus bel_2) \oplus bel_3)..... \oplus bel_n = \bigoplus_{i=1}^{n} bel_i \tag{3.5}$$

Based on the above, the outcome of combining the beliefs of Jim, Kim and Lim would be: $m(\theta) = 0.73$, $m(\overline{\theta}) = 0.27$ and $m(\Theta) = 0$, please note the influence of Jim's belief on the combination.

The D-S theory can be applied to the problem of combining the classification results of different classifiers by considering the evidence of each classifier as a BBA. Since the classifiers' evidence plays a crucial role in the combination performance, there is an increased interest in the proper estimation of such evidence. In the next section, we discuss how a number of existing classifier combination methods estimate the evidence of classifiers.

3.4 Existing Methods for Estimating the Evidence

Mandler and Schurmann [40] proposed a method that transformed distance measures of the different classifiers into evidence. This was achieved by first calculating a distance between learning datasets and a number of reference points in order to

estimate statistical distributions of intra- and interclass distances. A distance within a specific class label is called intraclass distance, while interclass distance is the distance between different classes. For both, the a posteriori probability function was estimated, indicating degree at which an input pattern belongs to a certain reference point. Then, for each class label, the class conditional probabilities were combined into an evidence value ranging between 0 and 1, which was considered as the BBA of that class. Finally, Dempster's combination rule was used to combine the BBAs of the different classifiers to give the final result. This approach departs from the traditional Bayesian method in the way the basic belief is assigned, here the distance measure is used to develop certain statistical model, where as in the Bayesian we pre-assume some sort of distribution over the data. As explained in [64], this method brought forward questions about the choice of reference vectors and the distance measure. Moreover, approximations associated with estimation of parameters of statistical models for intra- and interclass distances can lead to inaccurate measurements of the evidence.

In [42], $K+1$ classes were used to perform the classification task, where the $K^{th}+1$ class denotes that the classifier has no idea about which class the input comes from. For each classifier c^n, $n = 1.....N$, recognition, substitution and rejection rates (ϵ_r^n,ϵ_s^n and $1-\epsilon_r^n-\epsilon_s^n$) were used as a measure of BBA. Recognition rate is the accuracy of correctly classifying a test pattern, substitution rate is the misclassification rate and the rejection rate is a measure of the amount of patterns designated the class

label $K+1$.

Based on the above assumption, the algorithm in [42] is summarized as:

1. If the maximum output of a specific classifier belongs to $K+1$, then m_n has only one focal element Θ with $m_n(\Theta) = 1$.

2. When the maximum output belongs to one of the K classes, m_n has two focal elements θ_k and $\overline{\theta_k}$ with $m_n(\theta_k) = \epsilon_r^n$, $m_n(\overline{\theta_k}) = \epsilon_s^n$. As the classifier says nothing about any other proposition, $m_n(\Theta) = 1 - m_n(\theta_k) - m_n(\overline{\theta_k})$.

The drawback of this method is again the way the evidence is measured. There are two problems associated with this method. Firstly, many classifiers do not produce binary outputs, but rather probability like outputs. So, for the first case, it would be inaccurate assigning 0 to both $m_n(\theta_k)$ and $m_n(\overline{\theta_k})$. Secondly, this way of measuring evidence ignores the fact that classifiers normally do not have the same performance with different classes. This had a clear impact on the performance of this combination method when compared with other conventional methods especially the Bayesian approach [42].

Rogova [64] used several proximity measures between a reference vector and a classifier's output vector. The proximity measure that gives the highest classification accuracy was later transformed into evidences. The reference vector used was the mean vector, μ_k^n, of the output set of each classifier c^n and each class label k. A number of proximity measures, d_k^n , for μ_k^n and y^n were considered, y^n being the

output vector for the n^{th} classifier. For each classifier, the proximity measure of each class is transformed into the following BBAs:

$$m_k(\theta_k) = d_k^n,\ m_k(\Theta) = 1 - d_k^n \tag{3.6}$$

$$m_k(\overline{\theta_k}) = 1 - \prod_{l \neq k}(1 - d_l^n), m_k(\Theta) = \prod_{l \neq k}(1 - d_l^n) \tag{3.7}$$

The evidence of classifier c^n and class label k is obtained by combining the knowledge about θ_k. Finally, Dempster's combination rule was used to combine evidences for all classifiers to obtain a measure of confidence for each class label.

This was a promising idea. However, the major drawback is the way the reference vectors were calculated, where the mean of output vectors may not be the best choice. Also, trying several proximity measures and choosing the one that gives the highest classification accuracy is itself questionable and computationally expensive.

3.5 Chapter Summary

The Bayesian formulation of a given problem assumes a Boolean phenomenon which leads to *over-commitment* i.e the degree of belief we have in existence of certain hypothesis (say θ) has a causal effect on our belief in non-existence of the hypothesis ($\overline{\theta}$). Thus a small degree of belief in hypothesis θ automatically leads to large degree of belief to the negation of the hypothesis ($\overline{\theta}$). Thus our lack of knowledge leads to

a over-commited formulation of the problem. Dempster-Shafer theory of evidence (DST) in contrast to the Bayesian theory keeps as much belief in a hypothesis as implied by an *evidence*, there by avoiding the over-commitment. As such, under DST formulation of the problem, lack of belief does not mean disbelief, leading to an adequate representation of uncertainty. This ability of DST to represent uncertainty has attracted researchers to implement it in decision making problems where there is a lack of knowledge. Thus classifier combination approaches have been proposed using the DST formulation of the problem. After presenting the details of DST in combining classifiers, we propose in the next chapter a text-dependent speaker identification system using the DST framework.

Chapter 4

The DST Fusion of Homogeneous Distance Classifiers

4.1 Introduction

Accessing restricted areas or resources is becoming a regular part of our lives, whether we are trying to access a building or our bank accounts, we need some sort of identification or authentication. There are many ways to achieve this, for example an identity card, smart card etc, these all approaches however fall in the category of *"what you have?"*. However these means of authentication could be forged or stolen, thus the challenge is to use a biometric. Biometric rely on *"what we are?"* rather than *"what we have?"* to develop more secure and safe authentication systems. Among the biometric traits available for the purpose of person identification,

speech makes the most natural and obvious choice. Automatic speaker recognition (ASR) systems identify people utilizing the utterances.

The area of automatic speaker recognition could be subdivided into two specific branches:

1. Automatic Speaker Identification (ASI): This is a problem of recognizing that *who is talking?* In other words, the speaker desiring authentication, provides a test sample to the system. The system is then required to figure out who is the speaker among the N existing speakers in the database.

2. Automatic Speaker Verification (ASV): In this scenario, the user claims a certain identity (*I am Mr.X*) and the system verifies, (*Mr.X authenticated*) or rejects (*Access denied*) the user. Thus, verification is a 1 to 1 classification problem resulting into a binary outcome (authentication or rejection).

Depending upon the nature of the application, speaker identification or speaker verification systems, could be modeled to operate either in *text-dependent* or *text-independent* modes. For text-dependent ASR, the user is required to utter a specific password, while for text-independent ASR, there is no need for such a constraint. Success in both cases depends on the modeling of speech characteristics which distinguish one user from the other. Text-dependent ASR is used for applications where the user is willing to cooperate by memorizing the phrase or password to be spoken.

Research in the field of speaker recognition traces back to the early 1960s when

Lawrence Kersta at Bell Labs made the first major step in speaker verification by computers, where he introduced the term voiceprint for a spectrogram, which was generated by a complicated electro-mechanical device [73]. Since then, there has been a tremendous amount of research in the area. Starting from spectrogram comparisons, passing through simple template matching, dynamic-time wrapping, to more sophisticated statistical approaches like Gaussian Mixture Model (GMM) [74, 75, 76], Hidden Markov Model (HMM) [77, 78, 79], neural networks [80, 81], etc. We continue to witness new techniques on regular basis, however, despite robustness and reliability, speaker recognition systems, as all other biometric systems, have their own limitations. Furthermore, one classification method good for one application might not suit a different application. These observations have attracted the interest of researchers in trying to combine decisions from multiple classifiers to reach better recognition rates.

In this work, we propose to apply the Dempster-Shafer theory of evidence to the problem of speaker recognition. This chapter is organized as follows: In section 2, we propose our algorithm called **NNEF** (Nearest Neighbor based Evidence Fusion), followed by section 3 which discusses the two individual speaker recognition systems considered here. The chapter is concluded in section 4 with a number of experimental results.

4.2 Dempster-Shafer Formulation of the Problem

Consider the case of N classifiers denoted by $e^{(n)}$, where $n = 1, 2.....N$. Let $\mathbf{X}_k$ be the training data matrix for each class, $k = 1, 2, ...K$, K being total number of classes. We will assume here equal amount of training for each of the classes. Also let θ_k be the label for each class k. Now, the feature extraction module of each classifier extracts a feature matrix $\mathbf{X}_k^{(n)}$. We define a modeling function $\Omega(.)$ which models each class so that

$$\Omega(\mathbf{X}_k^{(n)}) = \mathbf{U}_k^{(n)}; \quad k = 1, 2, \ldots, K \tag{4.1}$$

$$n = 1, 2, \ldots, N \tag{4.2}$$

Let $\mathbf{z}$ be an input test pattern which is modeled in a similar way :

$$\Omega(\mathbf{z}) = \mathbf{Z} \tag{4.3}$$

For the case of a single classifier, the classification task is to assign class i to pattern $\mathbf{z}$ if:

$$D(\mathbf{U}_i, \mathbf{Z}) < D(\mathbf{U}_k, \mathbf{Z}) \; \forall \; k = 1, 2, ..., K \;\; (k \neq i) \tag{4.4}$$

where $\mathbf{U}_k$ is the model for each class k, and $\mathbf{U}_i$ being the nearest neighbor to $\mathbf{Z}$. $D(.)$ is a distance measure between the test pattern model ($\mathbf{Z}$) and the training pattern models for each class ($\mathbf{U}_k \quad k = 1, 2, \ldots, K$).

Assume now that we have N classifiers, so that each classifier operates on the test model independently to reach an independent decision.

Since for each classifier, the function $\Omega(.)$ models the patterns in the same manner, we propose the nearest neighbor distance $\underbrace{\min}_{k}^{(n)}\{D(\mathbf{U}_k^{(n)}, \mathbf{Z})\}$ as the evidence of our belief in the decision made by classifier n. Thus, the belief becomes a decreasing function (say $\psi(.)$) of this distance:

$$m^{(n)}(i) = \psi(\underbrace{\min}_{k}^{(n)}\{D(\mathbf{U}_k, \mathbf{Z})\}) \tag{4.5}$$

where $m^{(n)}(i)$ is our belief in classifier n for classifying $\mathbf{z}$ as class i.

One candidate for the function $\psi(.)$ could be the exponential function:

$$m^{(n)}(i) = \exp(-(\underbrace{\min}_{k}^{(n)}\{D(\mathbf{U}_k, \mathbf{Z})\})) \tag{4.6}$$

Hence the smaller the nearest neighbor distance measure, the greater is our belief in the decision of the classifier. In summary our algorithm works as follows:

1. Each class is modeled using the training data matrix $\mathbf{X}_k, k = 1, 2, \ldots, K$ and the function $\Omega(\mathbf{X}_k^{(n)}) = \mathbf{U}_k^{(n)}$.

2. Input test pattern $\mathbf{z}$ is also modeled using the same modeling function $\Omega(.)$, i.e $\Omega(\mathbf{z}) = \mathbf{Z}$.

3. A distance measure, $D(.)$ is then used to evaluate the distance between $\mathbf{Z}$ and each of the models $\mathbf{U}_k^{(n)}, k = 1, 2, \ldots, K$.

4. For each classifier, a label is given to the test pattern $\mathbf{z}$ which corresponds to minimum distance measure

$$d^{(n)} = \underbrace{\min}_{k}\{D(\mathbf{U}_k^{(n)}, \mathbf{Z}^{(n)})\} \tag{4.7}$$
$$n = 1, 2, \ldots, N$$
$$k = 1, 2, \ldots, K$$

5. We estimate our confidence in each classifier's decision as:

$$m^{(n)}(i) = \exp(-d^{(n)}) \tag{4.8}$$

6. We then combine all evidences $m^{(n)}$ $n = 1, 2, \ldots, N$ using Dempster-Shafer theory of evidence as follows:

$$m(k) = \frac{\sum_{j,l,\mathbf{A}_j \cap \mathbf{A}_l = k} m^{(1)}(\mathbf{A}_j) \ldots m^{(N)}(\mathbf{A}_l)}{1 - \sum_{j,l,\mathbf{A}_j \cap \mathbf{A}_l = \phi} m^{(1)}(\mathbf{A}_j) \ldots m^{(N)}(\mathbf{A}_l)} \tag{4.9}$$
$$k = 1, 2, \ldots, K$$

7. Class label j is assigned to test pattern if

$$j = \underbrace{\max}_{k}\{m(k)\}; \quad k = 1, 2, \ldots K \tag{4.10}$$

Some special cases to be considered are:

(a) if all classifiers reject a pattern, the consensus decision will then be rejection and thus our belief will be given to the frame of discernment $m(\Theta) = 1$.

(b) if a subset of classifiers say M rejects a test pattern, then these classifiers will be excluded and the decision will be made on basis of remaining $(N - M)$ classifiers.

The flow chart of NNEF algorithm is shown in figure 4.1.

4.3 The Developed Speaker Recognition System

We have developed two different speaker recognition systems, the main difference between the two systems resides in the different features used. Specifically we used the LPCC (Linear Prediction Cepstral Coefficients) and MFCC (Mel Frequency Cepstral Coefficients) methods of feature extraction.

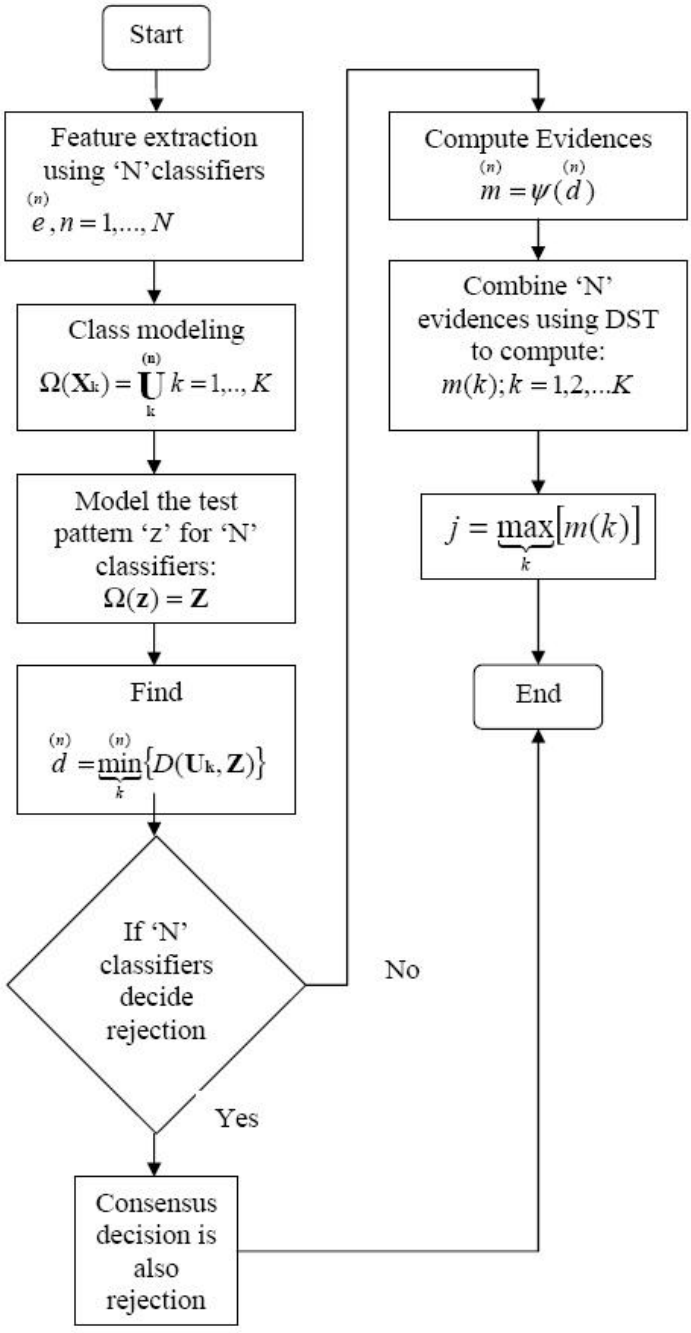

Figure 4.1: Flow chart for the proposed NNEF algorithm

4.3.1 Feature Extraction through LPCC

One of the most popular speech analysis techniques is that of linear prediction. Linear prediction analysis of speech has become the predominant technique for estimating the basic parameters of speech. Linear prediction analysis provides a good representation of speech characteristics at a low computational load.

The basic idea behind linear prediction analysis is that a specific speech sample at the current time can be approximated as a linear combination of past speech samples. Through minimizing the sum of squared differences (over a finite interval) between the actual speech samples and linearly predicted values a unique set of parameters or prediction coefficients can be determined. These coefficients form the basis for linear prediction analysis of speech. Thus a speech signal at time r can be approximated using p previous samples using an LP model of order p.

$$\widetilde{y}(r) = \sum_{j=1}^{p} a_j y(r-j) \tag{4.11}$$

where $\widetilde{y}(r)$ is the predicted speech sample, a_js are the prediction coefficients and $y(r-j)$s are the p previous speech samples. The error in prediction is given as:

$$e(r) = y(r) - \widetilde{y}(r) \tag{4.12}$$

The speech signal is first framed into blocks of approximately 30ms in length, the process is referred to as windowing. Fourier transform is then applied to each win-

dowed frame, to obtain the short-time spectrum. Power spectral density (PSD) is computed from the square of the magnitude of the spectrum. In the next step, IDFT (Inverse Discrete Fourier Transform) is applied to the PSD to obtain the autocorrelation function. Then the Levinson-Durbin recursion is used to estimate the LPC coefficients from the autocorrelation coefficients. Finally, the cepstral coefficients are found from the LPC coefficients in a recursive manner. The mathematical details could be found in [82]. We have used the work in [83] to implement an LPCC system.

4.3.2 Feature Extraction through MFCC

MFCC is perhaps the best known and the most popular feature extraction technique for speech signals. The main purpose of the MFCC is to imitate the behavior of a human ear. Psychophysical studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency f, measured in hertz, a subjective pitch is measured on a scale called the 'Mel' scale [84, 85, 86]. The Mel-frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000Hz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 Mels. Therefore we can use the following approximate formula to compute the Mels for a given frequency f in Hertz:

$$mel(f) = 2595 * \log_{10}(1 + \frac{f}{700}) \tag{4.13}$$

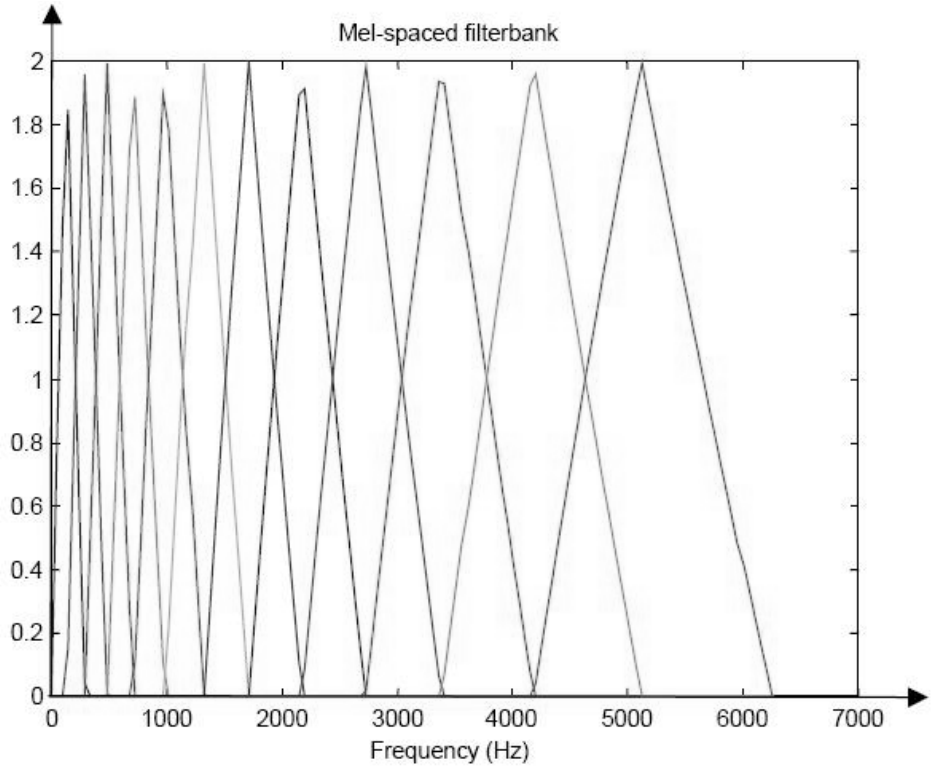

Figure 4.2: A typical Mel-spaced filter bank

One approach to simulate the subjective spectrum is to use a filter bank, spaced uniformly on the Mel scale (see figure 4.2). This filter bank is applied to the spectrum of the speech signal to get a Mel-spectrum. This Mel-spectrum when transformed back to time domain using the Discrete Cosine Transform (DCT) gives us the MFCC coefficients. Therefore if we denote the Mel power spectrum coefficients by S_k $k = 1, 2, ..., L$, k being the index of the Mel-spaced filters, then the

MFCC (c_n) are calculated as:

$$c_n = \sum_{k=1}^{K} (\log S_k) \cos(n(k - \frac{1}{2})\frac{\pi}{K}) \; n = 1, 2, \ldots, L \tag{4.14}$$

4.3.3 The Speaker Recognition System

We consider second order statistical modeling of speech, assuming a wide sense stationary process (WSS) as proposed in [87]. Let $\mathbf{X_k}$ be the training data matrix for class k, so that we have b samples available per class for training. Let $\{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_b}\}$, be a set of b feature vectors available for class k. Given the b patterns available for training per class, we model a class k through the mean and covariance matrix as follows:

$$\hat{\mathbf{x}} = \frac{1}{b} \sum_{i=1}^{b} \mathbf{x_i} \tag{4.15}$$

$$\tag{4.16}$$

$$\mathbf{U_k} = \frac{1}{b} \sum_{i=1}^{b} (\mathbf{x_i} - \hat{\mathbf{x}})(\mathbf{x_i} - \hat{\mathbf{x}})^t \tag{4.17}$$

Similarly for a test pattern $\mathbf{z}$, we derive a covariance matrix $\mathbf{Z}$. Once we have developed the second-order statistical models, we apply an *arithmetic-harmonic sphericity* measure [87] as the distance metric between $\mathbf{X}$ and $\mathbf{Z}$, thus

$$D_{sph}(\mathbf{U_k}, \mathbf{Z}) = \log \left[\frac{\mathrm{tr}(\mathbf{U_k}\mathbf{Z}^{-1})\mathrm{tr}(\mathbf{Z}{\mathbf{U_k}}^{-1})}{m^2} \right]; \;\; k = 1, 2, \ldots K. \tag{4.18}$$

where m is the dimension of the feature vector and tr($\mathbf{A}$) is trace of $\mathbf{A}$. The distance measures are mapped to $[0, 1]$ with a sigmoid function.

4.4 DST based Fusion of Speaker Recognition Systems using the Proposed NNEF Algorithm

We are now at the stage of testing our proposed fusion algorithm. Note that although the features are heterogeneous, they are reduced to the same distance metric, and thus we can safely take the distance as an evidence measure since there is no normalization issue. We have used a locally developed text-dependent database consisting of 40 classes. The password for authentication is the arabic greeting sentence *assalam-o-alaikum wa rahmatullah-e-wabarakathu*.

We tested our algorithm under three evaluation protocols:

1. **Evaluation Protocol 1**: Under the first evaluation protocol we verify our fusion algorithm (NNEF) when we assume one classifier is perfect. In our case it is the MFCC based classifier, which resulted in a 0% rejection and substitution (misclassification) rate. The NNEF algorithm also resulted in a 100% classification accuracy. Thus the evidence of MFCC classifier is strong enough to dominate the decision of LPCC, the DST fusion of the two thereby giving an optimal result.

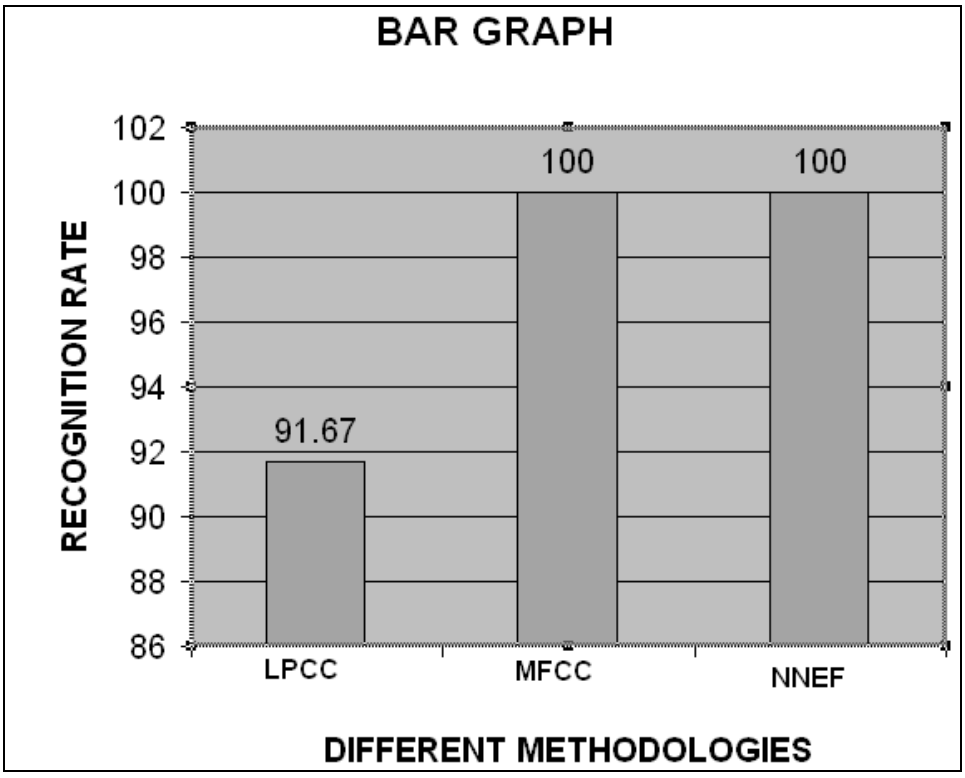

Figure 4.3: Bar graph representation of recognition rates for evaluation protocol 1

Method	Training	Testing	Classes	Recog.	Rej.	Subst.
MFCC	5	3	40	100%	0%	0%
LPCC	5	3	40	91.67%	0%	8.33%
NNEF	5	3	40	100%	0%	0%

Table 4.1: DST fusion results under evaluation protocol 1

Figure 4.3 and table 4.1 clearly show that when combining a perfect classifier with a poor classifier, the proposed NNEF algorithm opts for the perfect classifier, thereby avoiding the averaging process.

2. **Evaluation Protocol 2**: Under the second evaluation protocol, we modify our MFCC and LPCC classifiers by introducing a threshold value α for rejection. The values of α being 0.53 and 0.6 for MFCC and LPCC classifiers respectively.

Method	Training	Testing	Classes	Recog.	Rej.	Subst.
MFCC	5	3	40	91.67%	8.33%	0%
LPCC	5	3	40	87.5%	5%	7.5%
NNEF	5	3	40	95.83%	1.67%	2.5%

Table 4.2: DST fusion results under evaluation protocol 2

The recognition accuracy of the two classifiers has thus been reduced to 91.67% for MFCC and 87.5% for LPCC (see table 4.2 and figure 4.4). The DST based fusion of the two classifiers' decision using the proposed NNEF algorithm outperformed the two individual classifiers giving an improved recognition rate of 95.83%.

3. **Evaluation Protocol 3**: Under Evaluation Protocol 3, we make the problem

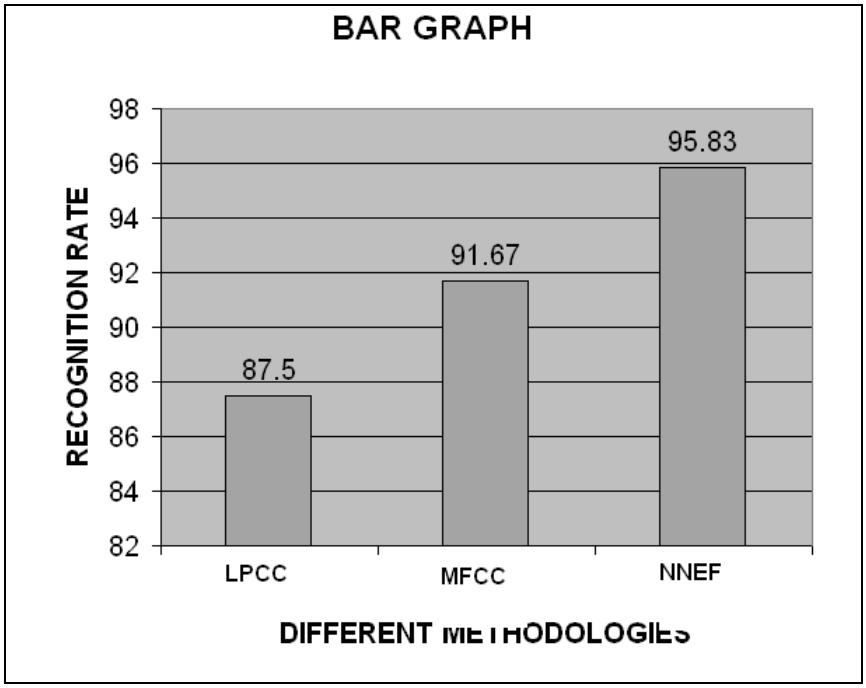

Figure 4.4: Bar graph representation of recognition rates for evaluation protocol 2

more complicated by adding white Gaussian noise to the speech data. The aim is to verify the robustness of the NNEF algorithm under noisy conditions.

Method	**Training**	**Testing**	**Classes**	**Recog.**	**Rej.**	**Subst.**
MFCC	5	3	40	89.17%	10.83%	0%
LPCC	5	3	40	90.83%	0%	9.17%
NNEF	5	3	40	96.67%	0%	3.33%

Table 4.3: Results of the NNEF algorithm for 30dB SNR

The results for noise contaminated speech data for different SNR values are shown in tables 4.3, 4.4 and 4.5 and figures 4.5, 4.6 and 4.7. For 30dB, 20dB and 15dB SNR the NNEF algorithm shows an improvement of 5.84%, 6.63% and 5% respectively, over the best of the combining classifiers.

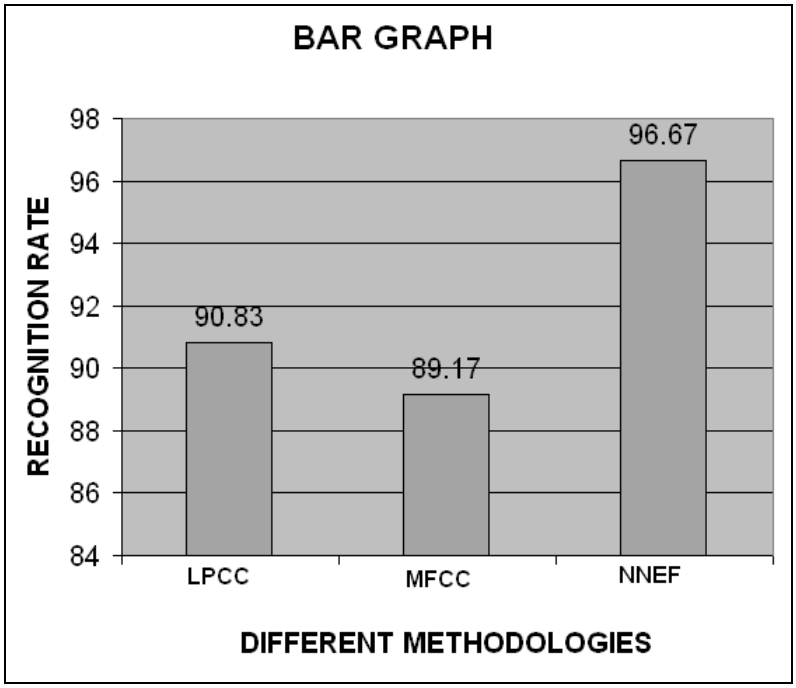

Figure 4.5: Bar graph representation of recognition rates for 30dB SNR

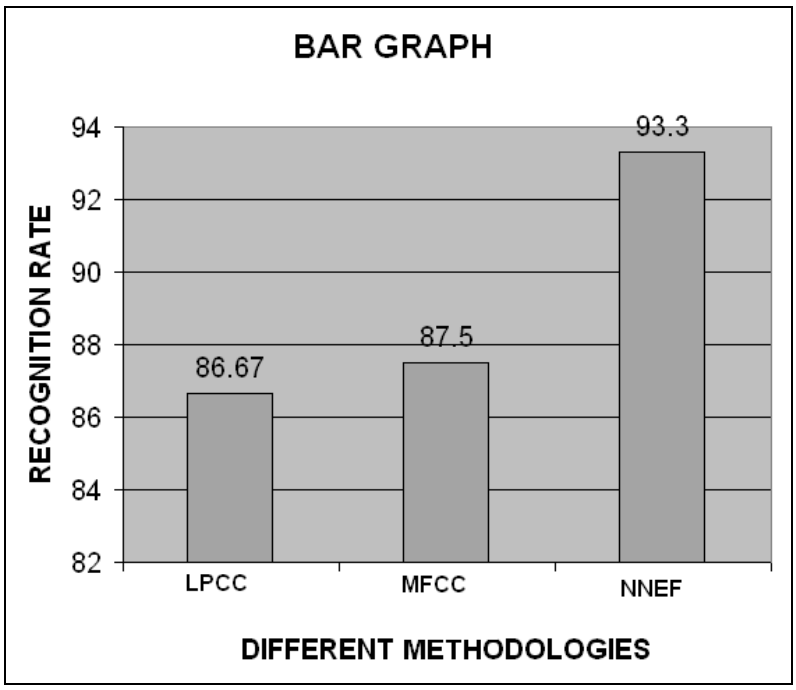

Figure 4.6: Bar graph representation of recognition rates for 20dB SNR

Method	Training	Testing	Classes	Recog.	Rej.	Subst.
MFCC	5	3	40	87.5%	12.5%	0%
LPCC	5	3	40	86.6%	2.5%	10.9%
NNEF	5	3	40	93.33%	0%	6.66%

Table 4.4: Results of the NNEF algorithm for 20dB SNR

Method	Training	Testing	Classes	Recog.	Rej.	Subst.
MFCC	5	3	40	85%	15%	0%
LPCC	5	3	40	83.3%	0%	6.66%
NNEF	5	3	40	90%	0%	10%

Table 4.5: Results of the NNEF algorithm for 15dB SNR

4.5 Chapter Summary

For homogeneous distance classifiers, the nearest neighbor (NN) distance is a strong evidence in favor of the decision made by the classifier. Based on this observation we proposed our algorithm called ***NNEF*** for fusion of different distance classifiers. The proposed algorithm has been tested on the speaker recognition problem and has shown to outperform the individual classifiers. The NNEF algorithm maintains its robustness even for speech data with the AWGN (Additive White Gaussian Noise). However the NNEF algorithm cannot combine heterogeneous classifiers. In the next chapter, we propose a DST fusion algorithm for multimodal biometrics (heterogeneous classifiers).

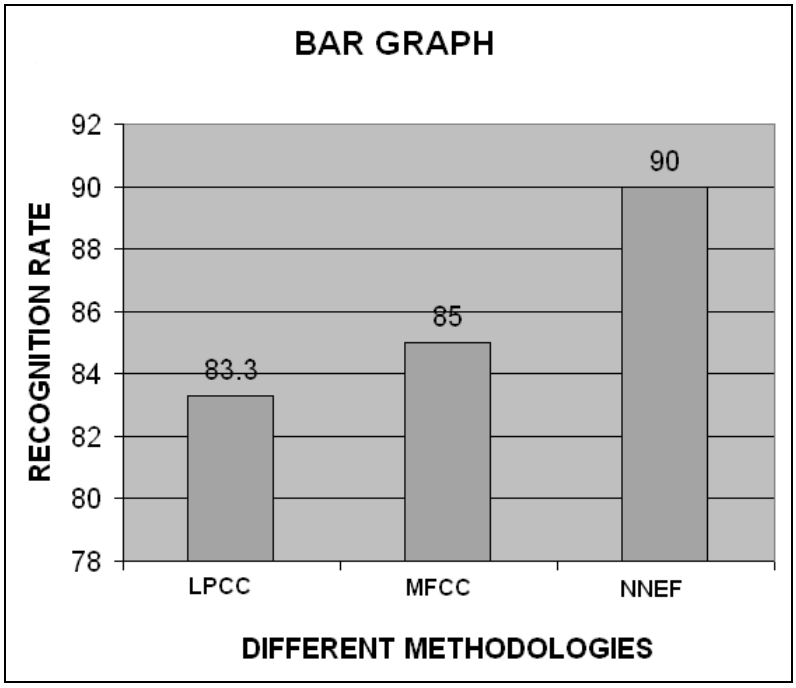

Figure 4.7: Bar graph representation of recognition rates for 15dB SNR

Chapter 5

The Proposed Multimodal Biometric Recognition System

A typical biometric system could be broadly divided into following four stages:

1. The sensor stage; collects the raw biometric data.

2. The feature extraction stage; converts the raw biometric data to a compact form (features).

3. The classification stage; which operates on the test pattern and the already developed client models to produce the decision variable.

4. The decision stage; performs the decision of either accepting a user or rejecting him.

In multimodal biometrics, the information fusion can occur at any of the above stages:

1. Combination at the data or feature level: Either the raw data or the features obtained from the data originating from multiple sources are fused.

2. Combination at the classification stage: The decision variables obtained from classifiers using different modalities are fused.

3. Combination at the decision stage: The decisions taken for different biometrics are combined to achieve an optimal decision.

5.1 The Multimodal Fusion Architecture

Depending upon the number of traits, sensors, and feature sets used, a variety of scenarios are possible in a multimodal biometric system, as shown in Figure 5.1.

1. Single biometric trait, multiple sensors: Multiple sensors record the same biometric trait. Thus, raw biometric data pertaining to different sensors are obtained. Chang et al. [88] acquired both 2D and 3D images of the face and combined these at the data level as well as the match score level to improve the performance of a face recognition system. Kumar et al. [89] described a hand-based verification system that combines the geometric features of the hand with palm prints at the feature and match score levels. Interestingly, in

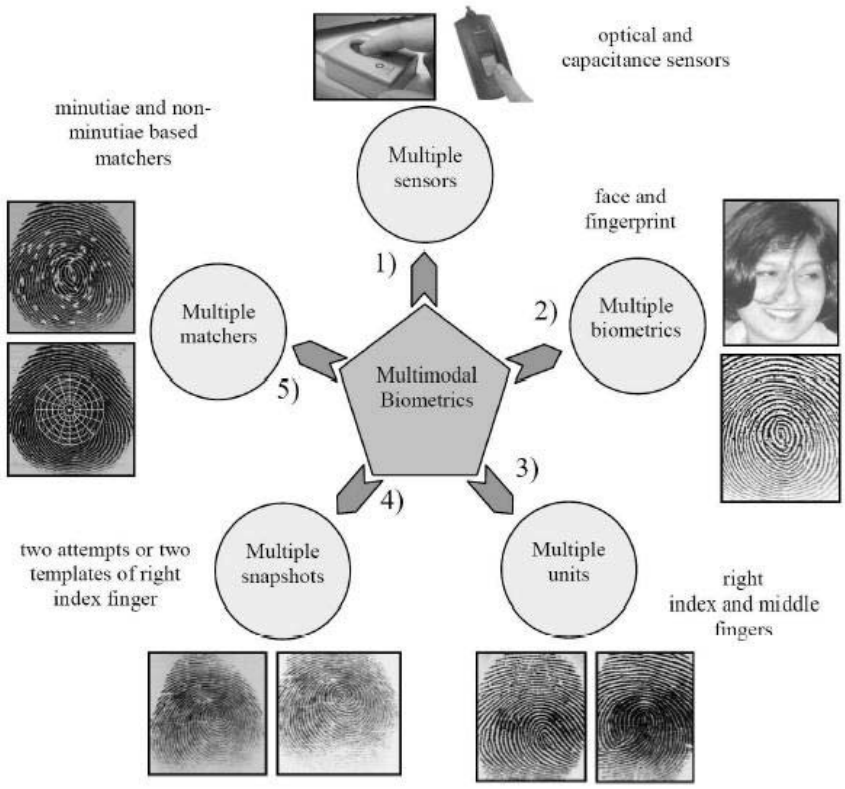

Figure 5.1: Different architectures of a multimodal biometric system

their experiments, fusion at the match score level results in better performance than fusion at the feature level. This could be due to the high-dimensionality of the fused feature set (the curse-of-dimensionality problem) and, therefore, the application of a feature reduction technique might have been more appropriate.

2. Single biometric trait, multiple classifiers: Unlike the previous scenario, only a single sensor is employed to obtain raw data; this data is then used by multiple classifiers. Each of these classifiers either operate on the same feature

set extracted from the data or generate their own feature sets. Jain et al. [90] used the logistic function to integrate the matching scores obtained from three different fingerprint matchers operating on the same minutiae sets. Ross et al. [91] combined the matching score of a minutiae-based fingerprint matcher with that of a texture-based matcher to improve matching performance. Lu et al. extracted three different types of feature sets from the face image of a subject (using PCA, LDA and ICA) and integrated the output of the corresponding classifiers at the match score level [92].

3. Single biometric trait, multiple units: In the case of fingerprints (or iris), it is possible to integrate information presented by 2 or more fingers (or both the irises) of a single user. This is an inexpensive way of improving system performance since this does not entail deploying multiple sensors nor incorporating additional feature extraction and/or matching modules.

4. Multiple biometric traits: Here, multiple biometric traits of an individual are used to establish the identity. Such systems employ multiple sensors to acquire data pertaining to different traits. The independence of the traits ensures that a significant improvement in performance is obtained. Brunelli et al. [93] used the face and voice traits of an individual for identification. A HyperBF network is used to combine the normalized scores of five different classifiers operating on the voice and face feature sets. Bigun et al. developed a statistical framework

based on Bayesian statistics to integrate speech (text dependent) and face data [94]. The estimated biases of each classifier is taken into account during the fusion process. Hong and Jain associated different confidence measures with the individual matchers when integrating face and fingerprint traits of a given user [95]. They also suggested an indexing mechanism wherein face information is used to retrieve a set of possible identities and the fingerprint information is then used to select a single identity. A commercial product called BioID [96] uses the voice, lip motion and face features of a user to verify identity.

In the next section we discuss the face and speaker recognition system we developed in this thesis.

5.2 The Proposed Multimodal Biometric System

5.2.1 The Face Recognition System

Face recognition, as a paradigm of a typical recognition system, has become a benchmark to the solutions of some of the main computer vision problems (invariance to view point; illumination change; occlusion; deformation due to changes of expression, age, make-up and hair style), as well to some of the main topics in statistical pattern recognition (feature selection; generalization; discriminability, etc.). This

is evident when the continuous publication of reviews and surveys is considered, from the earliest of Samal and Iyengar (1992), to the latest of Jain (2003), passing through the works of Valentin et al. (1994), Chellappa et al. (1995) and Fromherz (1998).

We have developed a PCA (Principal Component Analysis) based approach to the face recognition. The faces are mapped to PCA space, represented by eigen vectors, a minimum distance classifier is implemented for the decision purpose.

5.2.2 Mathematical Analysis of PCA

The mathematical description of the PCA has been presented in chapter 2 of this thesis.

5.2.3 The Speaker Recognition System

The speaker recognition system used is the same as the one developed in chapter 4 of this thesis.

5.3 A Dempster-Shafer Approach to Multimodal Biometrics

The Dempster-Shafer theory of evidence has proved its ability for adequate modeling of uncertainty in various applications. The problem of pattern classification is one

such application. However the theory has not yet been comprehensively explored in the area of person identification, specifically there is little research done in the implementation of the idea to the field of multimodal biometrics. The main hinderance being the multitude of heterogeneous classifiers that need to be combined. The term "heterogeneous" introduced in the context of classifiers, refers to the differences in classifiers to be combined. These may be:

1. The biometric traits to be combined are different.
2. The subspace techniques implemented for different classifiers are different.
3. In case of distance classifiers, the distance measure functions themselves are different

These deviations of different classifiers have been the main challenge for the purpose of belief estimation. Since for belief estimation we need such a parameter as an evidence which is same for the combining classifiers, we name such parameters as "global parameters". In this section we have proposed two global parameters to estimate our confidence in a classifier.

5.3.1 The Performance Parameters of a Classifier as the Evidence

The global parameter proposed as an evidence in committing belief in a classifier, has to be uniform and indifferent to the theory of the classifiers used. The performance

parameters of a classifiers such as recognition rate, substitution rate, rejection rate etc. are strong candidates for this purpose. This idea was adopted in [42] for the purpose of handwriting recognition. Asserting on the fact that the DST has not yet been used for multimodal biometric systems, we start our quest by forming basis on the approach discussed in [42]. We call our proposed algorithm as the RREF (Recognition Rate based Evidence Fusion).

Analytical Formulation

Let there be K number of classes $\theta_1, \theta_2, \ldots, \theta_K$, and let N be the total number of classifiers available $e^{(1)}, e^{(2)}, \ldots, e^{(N)}$. Let the $(K+1)th$ class denote the scenario when the classifier(s) has/have no idea about the class label of the input pattern and thus it corresponds to rejection.

Let us denote the recognition rate, substitution rate and rejection rate of a classifier "n" as $\epsilon_r^n, \epsilon_s^n$ and ϵ_{rej}^n respectively. Thus, the degree of confidence committed in the decision of classifier e^n is its recognition rate ϵ_r and the disbelief about the decision is its misclassification (substitution) rate. The rejection rate, corresponding to uncertainity, is distributed over the frame of discernment Θ.

Thus if classifier $e^{(n)}$ classifies a test pattern as θ_k, the DS-formulation of the

problem would be:

$$\left.\begin{array}{rcl} m_n(\theta_k) & = & \epsilon_r^n \\ m_n(\neg \quad \theta_k) & = & \epsilon_s^n \\ m_n(\Theta) & = & \epsilon_{rej}^n \end{array}\right\} \tag{5.1}$$

Thus for a given test pattern $\mathbf{z}$ we have N such BPAs $m_1(.), m_2(.), \ldots, m_N(.)$. Our problem is to combine all these evidences using Dempster's rule of combination:

$$m(.) = m_1(.) \bigoplus m_2(.) \bigoplus \ldots \bigoplus m_N(.) \tag{5.2}$$

Once a combined BPA is estimated for each class, the belief is calculated as given in equation 3.2,

$$\left.\begin{array}{l} bel(\theta_j) \\ bel(\neg \quad \theta_j) \end{array}\right\} ; j = 1, 2, \ldots, K \tag{5.3}$$

and thus a combined decision rule will be:

$$E(\mathbf{z}) = \underbrace{\max}_{j} [bel(\theta_j)] \tag{5.4}$$

Where $E(.)$ indicates the operation of the supervisor algorithm at the decision combination stage.

A few special cases are discussed below:

1. If all classifiers reject, a pattern the combined decision will also be rejection i.e $E(\mathbf{z}) = K + 1$

2. If a subset of classifiers $M < N$ rejects a pattern, such classifiers will be discarded in the decision making procedure, and thus we will have a total of $N - M$ evidences.

3. If there is a classifier with $\epsilon_r^n = 1$, it means that it always makes a correct decision. Thus, decision of such a classifier overrides the decision of other experts.

4. If there is a classifier with $\epsilon_s^n = 1$, i.e its decision is always wrong, such classifier will be discarded.

The flow chart of RREF algorithm is shown in figure 5.2

5.3.2 Experimental Evaluation of the RREF Algorithm

The RREF algorithm discussed above was tested on a 40 class problem, where the data for the face recognition part comes from the AT&T database (*www.uk.research.att.com/*) and the speaker database is the same used in chapter 4 of the thesis. First, we executed a validation procedure which would estimate the confidence in each classifier.

The validation procedure results clearly indicated that the recognition rate of the speech based classifier is better than that of the face classifier for the given thresholds

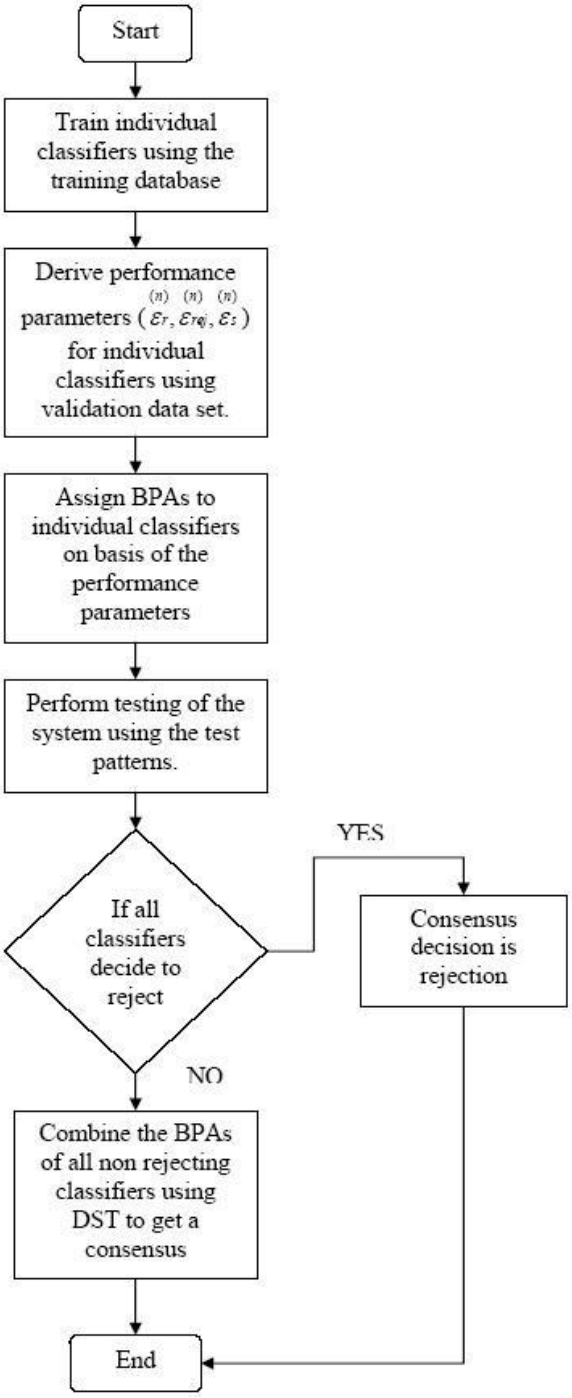

Figure 5.2: Flow chart for the RREF algorithm

Biometric	Training data	Validating data	Threshold	Recog.	Subst.	Rej.
Face	6	2	0.05	85%	2.5%	12.5%
Speech	6	2	0.57	87.5%	3.75%	8.75%

Table 5.1: Validation procedure for multimodal biometric system

Biometric	Training data	Testing data	Recog.	Subst.	Rej.
Face	6	2	86.25%	2.5%	11.25%
Speech	6	2	85%	8.75%	6.25%
RREF	6	2	90%	8.75%	1.25%

Table 5.2: Testing procedure for multimodal biometric system based on RREF algorithm

(refer to table 5.1). Based on these results we allocated the BPAs as $m_1 = 0.85$ and $m_2 = 0.875$. We then performed testing (please note that the test patterns are different from the validating patterns) with results displayed in table 5.2.

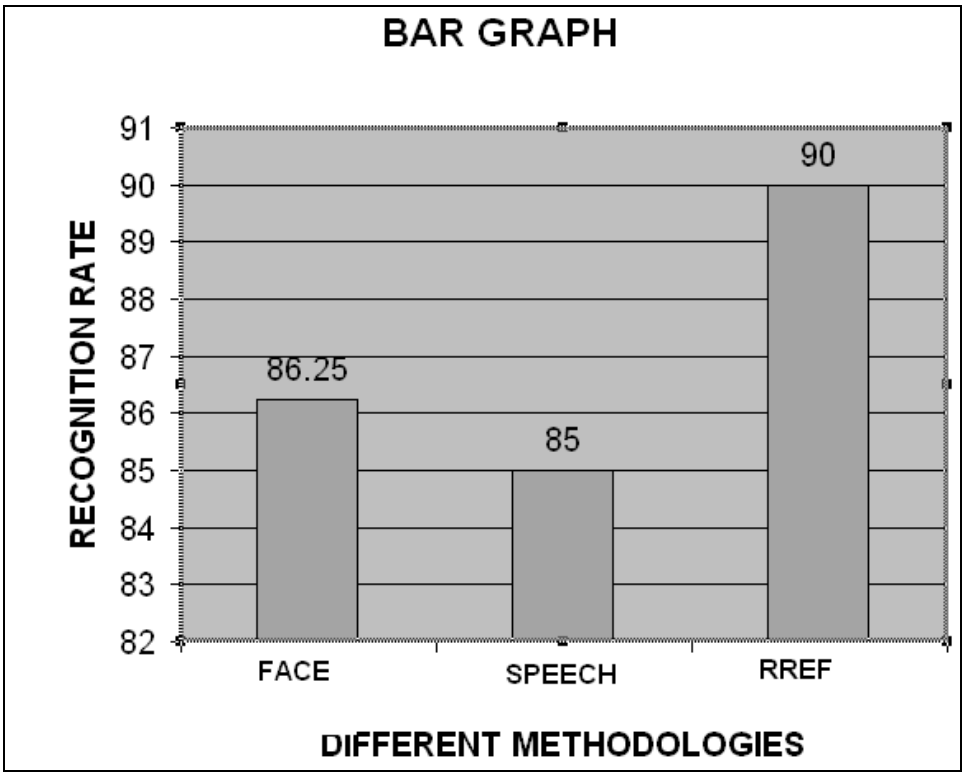

Figure 5.3: Bar graph representation of recognition rates for RREF algorithm

Figure 5.4: A subject of AT&T database with various poses.

The results clearly showed that the RREF algorithm outperforms the individual classifiers in recognition accuracy by almost 5%. Thus, the approach of belief estimation using the performance parameters as proposed in [42] for the handwriting recognition problem is also valid for the person identification problem. However there are a few comments to be made in this regard:

1. First, the system needs some validation data to estimate the confidence in the different classifier. If we don't have enough data, the validation procedure will not result in an adequate estimation of the belief.

2. In many cases the validation data is not a good representative of a particular subject, in this case, there will be an erroneous confidence estimation. Thus the confidence estimation procedure is subjective to the quality of the validating data available. As has been shown in the previous results, under validating protocol, the speech classifier seems to be a better classifier, while for the testing session, the face classifier outperformed the speech classifier. Thus, there is always an element of ambiguity involved in the RREF algorithm.

3. Also, using global recognition rate of a classifier as an evidence is itself not always true, since a classifier is likely to have different recognition rates for different classes. The same point is raised in [97], in which a class based recognition rate is used as the evidence for the problem of handwritten numerals (10 classes). Our criticism on this approach is two fold: firstly, this approach

makes the system too complex, secondly, for the person identification problem with large number of classes, the approach is impractical and burdensome.

5.3.3 The Statistical Measure of the Decision Variable as an Evidence

In the previous section we showed that there is no optimal approach to combine heterogeneous distance classifiers for the problem of person identification. In this section, we propose a new method based on the statistical measures of the participating experts. The proposed method is developed on a rational that the parameter to estimate evidence should be independent to the theories of the combining classifier. This would result in a supervisor algorithm which would be able to combine heterogeneous distance classifiers.

Different distance classifiers available diverge from each other in their theories, however no matter what the theory behind a certain distance classifier is, the main aim is to amplify the *inter class distances*. Thus, in general, it is quite adequate to say that the more a classifier is able to discriminate between the classes, the better the classification result is. Based on this observation, we propose to use the second order statistical measure of the decision variable (*inter class distances*) as an evidence of our belief. We call this algorithm as the **VEF** i.e Variance based Evidence Fusion.

5.3.4 Analytical Formulation of the VEF Algorithm

Let there be N classifiers $e^{(1)}, e^{(2)}, \ldots, e^{(N)}$ used for a K class problem, such that each class is denoted as θ_k, $k = 1, 2, \ldots, K$. Thus under Dempster-Shafer framework we have a frame of discernment denoted as: $\Theta = \{\theta_1, \theta_2, \ldots, \theta_K\}$.

Let $\mathbf{X}_k$ be the training data matrix for class θ_k, $k = 1, 2, \ldots, K$. This training data is used to develop a client model for class "k". Let the modeling function for classifier $e^{(n)}$ be $\Omega^{(n)}(.)$, $n = 1, 2, \ldots, N$, such that each modeling function is different from the other in its theory and approach, resulting in the heterogeneous models for each classifier.

$$\left.\begin{array}{rcl} \Omega^{(1)}(\mathbf{X}_k) & = & \mathbf{U}_k^{(1)} \\ \Omega^{(2)}(\mathbf{X}_k) & = & \mathbf{U}_k^{(2)} \\ \vdots & & \\ \Omega^{(N)}(\mathbf{X}_k) & = & \mathbf{U}_k^{(N)} \end{array}\right\} k = 1, 2, \ldots, K \tag{5.5}$$

Please note that $\mathbf{U}_k^{(1)}, \mathbf{U}_k^{(2)}, \ldots, \mathbf{U}_k^{(N)}$ are the client models developed by classifiers $e^{(1)}, e^{(2)}, \ldots, e^{(N)}$ respectively, for class θ_k, $k = 1, 2, \ldots, K$.

Now for a test pattern $\mathbf{z}$, the modeling operation results in:

$$\left.\begin{array}{rcl} \Omega^{(1)}(\mathbf{z}) & = & \mathbf{Y}^{(1)} \\ \Omega^{(2)}(\mathbf{z}) & = & \mathbf{Y}^{(2)} \\ \vdots & & \\ \Omega^{(N)}(\mathbf{z}) & = & \mathbf{Y}^{(N)} \end{array}\right\} \tag{5.6}$$

Now the classifier operation for classifier $e^{(n)}$ with test pattern model $\mathbf{Y}^{(n)}$ and training models for each class $\mathbf{U}_k^{(n)}, k = 1, 2, \ldots, K$ is reduced to be:

$$d_{min}^{(n)} = \underbrace{\min}_{k}[D^{(n)}(\mathbf{U}_k^{(n)}, \mathbf{Y}^{(n)})] \qquad k = 1, 2, \ldots, K \tag{5.7}$$

For classifier $e^{(n)}$, we define the inter-class distances as:

$$d_k^n = D^{(n)}(\mathbf{U}_k^{(n)}, \mathbf{Y}^{(n)}) \tag{5.8}$$

Note that these distances are normalized and mapped to [0,1]. For classifier e^n, we arrange the K inter class distances in a vector $\mathbf{d}^{(n)}$ so that:

$$\mathbf{d}^{(n)} = \begin{bmatrix} d_1^{(n)} \\ d_2^{(n)} \\ \vdots \\ d_K^{(n)} \end{bmatrix} \tag{5.9}$$

Perceiving the affiliation between performance of a classifier and these inter class distance measures, we argue that the more apart these distance are the better the classifier is. Thus if a distance classifier is able to perform clustering of classes more distinctively, it has a better chance of making correct decision. Harnessing upon this observation, we propose the second order statistical measure, i.e variance, of these inter class distances as an evidence in favor of the classifier's decision, thus

$$v^{(n)} = variance[\mathbf{d}^{(n)}] \tag{5.10}$$

Our evidence should be an increasing function of $v^{(n)}$, let $\psi(.)$ be this increasing function which maps variance to [0,1], thus if $e^{(n)}$ decides class θ_j for a test pattern $\mathbf{z}$ with a confidence measure $m^{(n)}(\theta_j)$ then

$$m^{(n)}(\theta_j) = \psi(v^{(n)}) \tag{5.11}$$

There could be many possible candidates for $\psi(.)$, however we have used sigmoid function (logsig) for our experiments, thus

$$m^{(n)}(\theta_j) = logsig(v^{(n)}) \tag{5.12}$$

$$m^{(n)}(\Theta) = 1 - m^{(n)}(\theta_j) \tag{5.13}$$

Thus using Dempster's rule of combination, we have combined our belief for the classes:

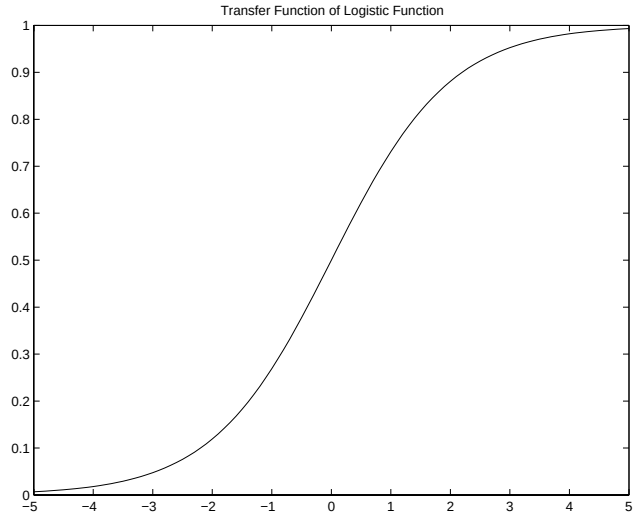

Figure 5.5: Transfer function of the logistic function

$$m(\theta_k) = \frac{\sum_{j,l,\mathbf{A}_j \cap \mathbf{B}_l = \theta_k} m^{(1)}(\mathbf{A}_j) \ldots m^{(N)}(\mathbf{B}_l)}{1 - \sum_{j,l,\mathbf{A}_j \cap \mathbf{B}_l = \phi} m^{(1)}(\mathbf{A}_j) \ldots m^{(N)}(\mathbf{B}_l)} \tag{5.14}$$

$$k = 1, 2, \ldots, K$$

Finally the decision rule is

$$\theta_j = \underbrace{\arg\max}_{k}[m(\theta_k)] \tag{5.15}$$

The flow chart of VEF algorithm is shown in figure 5.6.

5.3.5 Experimental Results for VEF Algorithm

We will now test the VEF algorithm for the heterogeneous classifiers case. To check the validity of the approach, we have carried out experiments under two evaluation

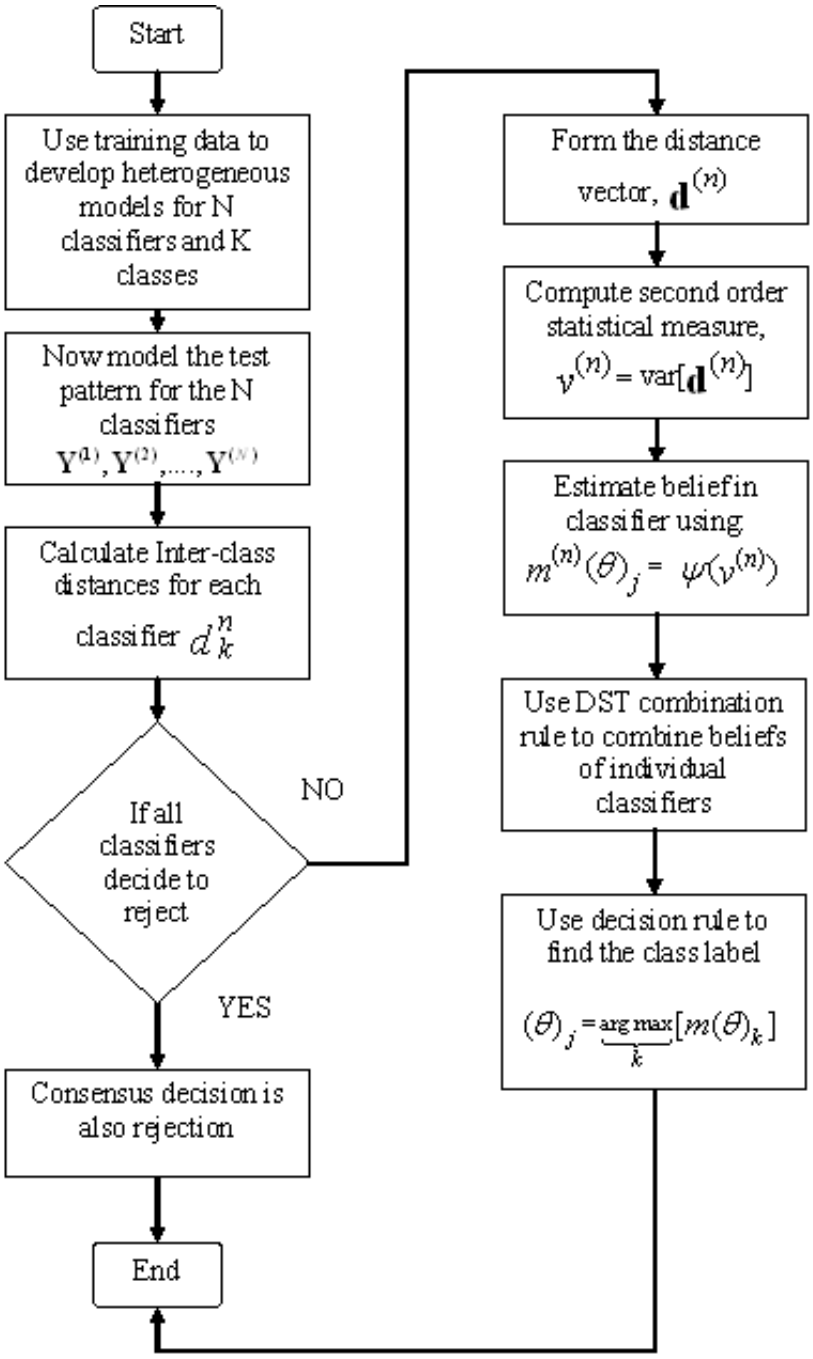

Figure 5.6: Flow chart for the VEF algorithm

protocols, both utilizing the benchmark databases.

Evaluation Protocol 1

Under evaluation protocol 1 we have tested the algorithm for a 15 class bench mark database called the YALE database (*www.cvc.yale.edu/projects/yalefaces/*). For the YALE database we have 15 subjects with 11 samples per subject. The face recognition system is trained for 7 and tested for 4 samples per subject with the value of threshold $\alpha = 0.03$. The speech recognition system is trained on 6 and tested on 4 samples per class with a threshold value of $\alpha = 0.57$. The speech database used is the same used in chapter 4 of the thesis. The fusion results of VEF algorithm are shown in table 5.3.

Biometric	**Recognition Rate**	**Substitution Rate**	**Rejection Rate**
Face	91.67%	3.33%	5%
Speech	85%	15%	0%
VEF	96.6%	3.33%	0%

Table 5.3: Classification results for a multimodal biometric system based on VEF algorithm using YALE database

The results of the VEF fusion show improved performance. The amalgamation algorithm has been able to improve the results of the individual classifiers, the recognition accuracy being better than the best of the combining classifiers by a substantial margin of 5%.

However a close observation of the results in table 5.3 shows that the face clas-

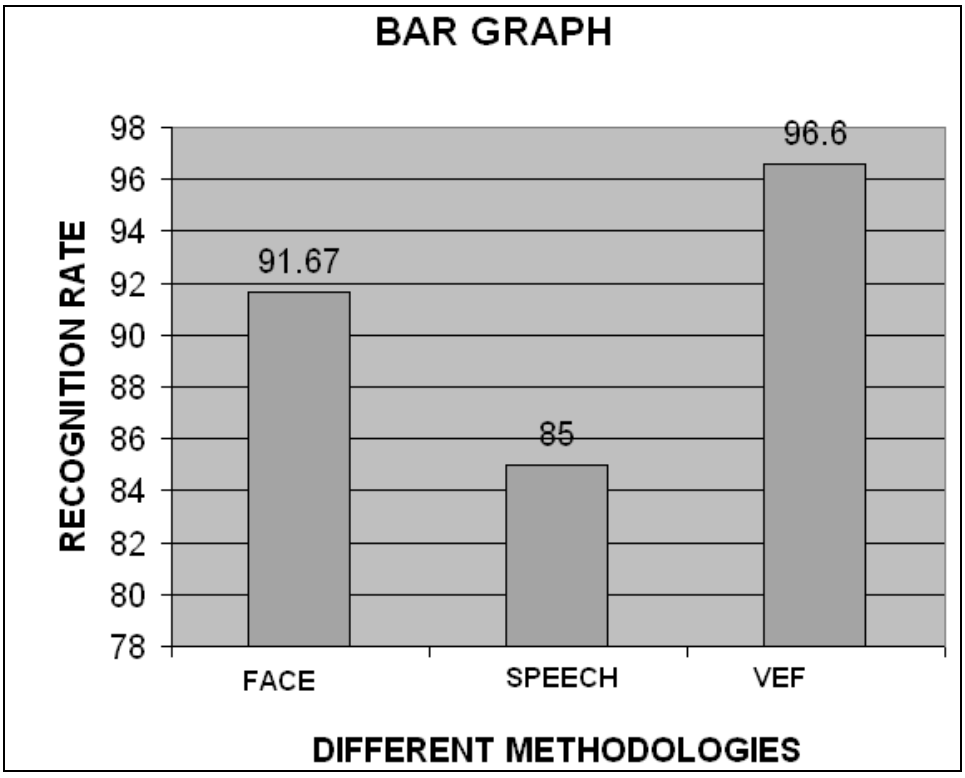

Figure 5.7: Bar graph representation of recognition rates for VEF algorithm

sifier offers 91.67% recognition with a rejection of 5%, where as the VEF algorithm actually gives 96.6% recognition with 0% rejection, which could mislead us to a conclusion that had the face classifier been made a 0% rejection classifier it would have given similar results as the VEF algorithm. To avoid this misleading conclusion, we performed experiments by removing the threshold for the face classifier and thus making the ambiguity very clear.

Biometric	**Recognition Rate**	**Substitution Rate**	**Rejection Rate**
Face	93.3%	6.7%	0%
Speech	85%	15%	0%
VEF	96.6%	3.33%	0%

Table 5.4: Comparison of performance between the face and the VEF classifier at 0 rejection.

The results in table 5.4 clearly show that even for the no rejection case the VEF algorithm outperforms the best of the combining classifiers. Thus the VEF algorithm is superior to both the speech and the face classifiers individually.

Evaluation Protocol 2

We continue our experiments for the VEF algorithm, under evaluation protocol 2 where we have increased the number of classes from 15 to 40 making the problem more complicated. The face database is the AT&T benchmark face database with 40 subjects and 10 samples per subject. The speech data base is the same used in chapter 4 of the thesis. The threshold values for face and speech are $\alpha = 0.05$ and $\alpha = 0.57$ respectively. The experiments conducted for different combinations of

Figure 5.8: A subject of the YALE database with different poses and illumination conditions.

training and testing data show dominance of the VEF algorithm.

The results of the experiments are shown in table 5.5 which discernibly justify our rationale of adopting the clustering ability of a classifier as an evidence. Extensive experiments were carried out for different combination of training and testing data and the VEF algorithm has shown its dominance in all cases.

However a similar argument can be made here as was made in Evaluation Protocol 1, meaning that the rejection rate of the face classifier is too high, and had the face classifier been a no rejection classifier, its performance would have been comparable to that of the VEF algorithm. To avoid this misleading conclusion, we made the face classifier a non rejecting classifier so that a fair comparison of the

Biometric	Training data	Testing data	Recog.	Subst.	Rej.
Face	6	2	86.25%	2.5%	11.25%
Speech	6	2	90%	10%	0%
VEF	6	2	95%	5%	0%
Biometric	**Training data**	**Testing data**	**Recog.**	**Subst.**	**Rej.**
Face	7	3	89.17%	2.5%	8.33%
Speech	7	3	90%	10%	0%
VEF	7	3	93.3%	6.7%	0%
Biometric	**Training data**	**Testing data**	**Recog.**	**Subst.**	**Rej.**
Face	6	4	85.62%	2.5%	11.87%
Speech	6	4	91.87%	8.13%	0%
VEF	6	4	95%	5%	0%
Biometric	**Training data**	**Testing data**	**Recog.**	**Subst.**	**Rej.**
Face	5	5	81%	8.5%	10.5%
Speech	5	5	89.5%	10.5%	0%
VEF	5	5	90.5%	9.55%	0%

Table 5.5: Classification results for a multimodal biometric system based on VEF algorithm using AT&T database

performance could be made.

Biometric	Training data	Testing data	Recog.	Subst.	Rej.
Face	5	5	84%	16%	0%
Speech	5	5	89.5%	10.5%	0%
VEF	5	5	90.5%	9.5%	0%

Table 5.6: Comparison of the face classifier with the VEF algorithm for no rejection condition

The results shown in 5.6 clearly demonstrate that making the face classifier a non rejecting classifier actually increases the substitution rate, which is undesirable, rather than contributing towards the recognition rate.

We assert the fact that the proposed algorithm is independent of the theories

behind the combining classifiers, and thus any number of heterogeneous distance classifiers could be combined using this algorithm. The VEF algorithm is superior to RREF algorithm because of the following reasons:

1. Recognition rate (ϵ_r) in RREF algorithm used as an evidence is highly dependent upon the quality and quantity of the validation data while the VEF algorithm uses the statistical measures of the decision variables as evidence.

2. There has to be a validation procedure for RREF algorithm, which most of the time is considered to be a burden on the system. The VEF algorithm does not need such procedure, thereby making system more robust and simple.

3. In RREF algorithm we ignored the fact that for different classes a classifier may have different recognition rates, thus use of the global recognition rate solely as an evidence is somehow unjust.

5.4 Chapter Summary

The Dempster-Shafer theory of evidence has not yet been utilized for the multimodal biometric recognition problem. In this chapter, we have proposed two algorithms to achieve the DST fusion of face and speech traits. The RREF algorithm makes use of the performance parameters of the individual classifiers for estimating belief, whereas the VEF algorithm utilizes the second order statistics of the decision variable to

commit belief. The fusion algorithms have been shown to outperform the individual classifiers with respect to recognition accuracy.

Chapter 6

Conclusions

In this chapter, we give a summary of the thesis and major findings of the research performed. We then describe future research directions that may be conducted in the area of classifiers' fusion using the Dempster-Shafer theory of evidence.

6.1 Thesis Summary

It is well known that different classifiers give complementary information for a specific pattern recognition problem. Thus, combining classifiers efficiently promises an improved recognition accuracy which is better than the best of the combining classifiers. The Dempster-Shafer theory of evidence, due to its ability of adequate modeling of uncertainty, has been proposed for combining different classifiers and has shown to improve the overall recognition accuracy. However the area of person

identification using biometrics has not yet been explored in this regard, specifically there have been no research carried for combining multimodal biometrics using the DST. Perceiving the ability of the DST under conditions of lack of knowledge, we proposed the use of the theory for the problem of person identification using biometrics.

In **chapter 4** of the thesis we have proposed the **NNEF** algorithm for combining *homogeneous* distance classifiers. The term "homogeneous" introduced in this context refers to the ability of different classifiers to reduce input patterns to uniform distance measures. The NNEF algorithm conceives the fact that for distance classifiers, the nearest neighbor distance is a strong evidence to commit ones belief in the decision of the classifier. Thus, under such framework, the NNEF algorithm converts the nearest neighbor distances into evidences, the DST is used to fuse these evidences to get a consensus decision. The NNEF algorithm was implemented for the speaker recognition problem. Extensive experiments were carried out using a 40 class, text dependent speech data base, the NNEF algorithm was shown to outperform the individual classifiers even in the presence of AWGN.

However, this algorithm is not applicable for a multimodal biometric recognition system. There is no implementation of the DST sofar for the problem of multimodal biometric systems. We have proposed the **RREF** algorithm in **chapter 5** of the thesis. The RREF algorithm apprehend the fact that the performance parameter (recognition rate) of a classifier is a good candidate to be used as an evidence in

favor of the classifier's decision. Thus, we proposed the RREF algorithm which, in the validation stage, works on the validation database to evaluate performance parameter of the classifier, which is then converted into an evidence. At the testing stage, these evidences are combined using the DST theory to get a consensus decision. The algorithm was implemented in fusing face and speech data in a 40 class problem. The experimental results showed the dominance of the RREF algorithm over the individual classifiers.

However the RREF algorithm had a few drawbacks:

1. First of all, the system needs some validation data to estimate the confidence in the classifiers. If we don't have enough data, the validation procedure will not result in an adequate estimation of the belief.

2. In many cases, the validation data is not a good representative of a particular subject, in this case, there will be an erroneous confidence estimation. Thus, the confidence estimation procedure is subjective to the quality of the validating data available.

3. Also, using the global recognition rate of a classifier alone as an evidence is itself not very accurate, since a classifier is likely to have different recognition rates for different classes. The same point is raised in [97], in which a class based recognition rate is used as the evidence for the problem of handwritten numerals (10 classes).

Based on the above, there was a need of a technique which can dynamically estimate the belief in a given classifier.

This is a common observation that for the case of distance classifiers the performance depends on the clustering ability of the classifier. Which means that if a classifier is able to cluster classes more distinctly, it is a better classifier than the one with poor clustering ability. As such, we proposed the second order moment of the inter-class distances to be an evidence in favor of the classifier. This observation leads to the formulation of the **VEF** algorithm proposed in **chapter 5** of the thesis. For the first time, statistical measures of classifiers are used to estimate evidence. We implemented the VEF algorithm for fusing decisons from face and speech classifiers. Extensive experiments were carried out using standard databases, which confirm our argument on the power and advantages of this VEF algorithm.

6.2 Recommendations for the Future Research

Our recommendations for the future research include:

1. The RREF and VEF algorithms developed here, have shown their success the in case of fusion of face and speech classifiers. We strongly recommend the use of these algorithms for different biometric traits, for example iris, finger prints, hand geometry etc

2. We propose the use of these algorithms to fuse classifiers which perform data decomposition using wavelets prior going into the feature space.

3. We recommend the investigation of high order statistical measures for belief estimation.

6.3 Conclusion

In this thesis we investigated the problem of classifier combination under the framework of the theory of evidence. We have developed three novel techniques for the fusion of both unimodal and multimodal biometric recognition systems:

1. *The NNEF Algorithm* is developed to fuse the homogeneous classifiers. It estimates the belief in a classifier using the nearest neighbor distance.

2. *The RREF Algorithm* utilizes the performance parameters of a classifier such as recognition rate, misclassification rate and rejection rate to estimate belief.

3. *The VEF Algorithm* utilizes the second order moment of the decision variables to estimate belief in a classifier's performance.

We performed extensive experiments to verify the validity of the developed algorithms. The NNEF algorithm was implemented for the speaker recognition problem. It has shown to outperform the individual experts. The RREF and the VEF algo-

rithms were tested for the fusion of audio and visual information. Both algorithms outperformed the individual classifiers.

Bibliography

[1] J. A. Prabhakar.S, Pankanti.S, "Biometric recognition: security and privacy concerns," *Security and Privacy Magazine, IEEE*, vol. 1, pp. 33 – 42, Mar-Apr 2003.

[2] S. Haykin, "Neural network a comprehensive foundation," *Prentice Hall*, 1999.

[3] R.Schalkoff, "Pattern recognition: statistical structural and neural approaches," *Wiley*, 1992.

[4] R. J.Kittler, M.Hatef and J.Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 226–239, 1998.

[5] D. M. M. Golfarelli, D. Maio, "On the error-reject tradeoff in biometric verification systems," *IEEE Trans. on Patt. Anal. and Mach. Intell*, vol. 19, pp. 786–796, 1997.

[6] A. Ross and A. K. Jain, "Information fusion in biometrics," *Pattern Recognition Letters*, vol. 24, pp. 2115–2125, 2003.

[7] C. A. S. R. P. W. D. L. I. Kuncheva, C. J. Whitaker, "Is independence good for combining classifiers?," *Proc. of Intl Conf. on Pattern Recognition (ICPR)*, vol. 2, pp. 168–171, 2000.

[8] I. Naseem and M. Deriche, "Human face detection in complex color images," *Third IEEE International Conference on Systems, Signals and Devices, SSD'05*, 2005.

[9] A. J.Cai and C.Yu, "Detecting human faces in color images," *Wright State University, U. of Illinois*.

[10] G.Wyszecki and W.s.Styles, *Color Science: Concepts and Methods, Quantitative Data and Formulae*. John Wiley and Sons, New York, 2nd ed., 1982.

[11] Y.Gong and M.Sakauchi, "Detection of regions matching specified chromatic features," *Computer Vision and Image Understanding*, vol. 61, no. 2, pp. 263–269, 1995.

[12] J. Yand and A. Waibel, "A real-time face tracker," *CMU CS Technical Report*.

[13] P. S. F. Kovac, J.; Peer, "Illumination independent color-based face detection," *Image and Signal Processing and Analysis*, vol. 1, September 2003.

[14] D. K. W. Son Lam Phung; Bouzerdoum, A.; Chai, "A color-based approach to automatic face detection," *3rd IEEE International Symposium on Signal Processing and Information Technology*, December 2003.

[15] V. Ming-Jung Seow; Valaparla, D.; Asari, "Neural network based skin color model for face detection," *Applied Imagery Pattern Recognition Workshop*, 2003.

[16] W. C. Chellappa, R. and S. Sirohey, "Human and machine recognition of faces:a survey," *Proceedings of the IEEE*.

[17] M. Grudin, "On internal representations in face recognition systems," *Pattern Recognition*.

[18] A. MSamal and P. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: A survey," *Pattern Recognition*.

[19] G. Cottrell, "Connectionist models of face processing:a survey," *Pattern Recognition*.

[20] B. Moghaddam, "Principal manifolds and bayesian subspaces forvisual recognition," *ICCV*, 1999.

[21] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 696–710.

[22] O. M. El-Khamy, S.E.Abdel-Alim, "Neural network face recognition using statistical feature and skin textureparameters," *Radio Science Conference NRSC*, 2001.

[23] F. S. V. E. Bouattour, H., "Neural nets for human face recognition," *Neural networks IJCNN.*, 1992.

[24] S. R.A.Jacobs and G.E.Hinton, "Adaptive mixtures of local experts," *Neural Computation*, vol. 3, pp. 79–87, 1991.

[25] M.Jordan and R.Jacobs, "Hierarchial mixtures of experts and the em algorithm," *Neural Computation*, vol. 6, pp. 181–214, 1994.

[26] M. R.A.Jacobs and A.G.Barto, "Task decomposition through competition in a modular connectionist architecture- the what and where vision tasks," *Cognitive Science*, vol. 15, pp. 219–250, 1991.

[27] M. A. J. Cao and M. Shridhar, "Recognition of handwritten numerals with multiple feature and multistage classifier," *Pattern Recognition*, vol. 28, pp. 153–160, 1995.

[28] J. Zhou and T. Pavlidis, "Discrimination of characters by a multi-stage recognition process," *Pattern Recognition*, vol. 27, pp. 1539–1549, 1994.

[29] M. E.-R. H. El-Shishini, M.S. Abdel-Mottaleb and A. Shoukry, "A multi-stage algorithm for fast classification of patterns," *Pattern Recognition Letters*, vol. 10, pp. 211–215, 1989.

[30] M. Kurzynski, "On the identity of optimal strategies for multistageclassifier," *Pattern Recognition Letters*, vol. 10, pp. 39–46, 1989.

[31] F. Kimura and M. Shirdhar, "Handwritten numerical recognition based on multiple algorithms," *Pattern Recognition*, vol. 24, pp. 969–983, 1991.

[32] H. L. C. Tung and J. Tsai, "Multi-stage pre-candidate selection in handwritten chinese character recognition systems," *Pattern Recognition*, vol. 27, pp. 1093–1102, 1994.

[33] M. Fairhurst and A. Rahman, "Generalised approach to the recognition of structurally similar handwritten characters using multiple classifiers," *IEE Proceedings-Vision, Image and Signal Processing*, vol. 144, pp. 15–22, 1997.

[34] J. W. K. Huang and H. Yan, "Off-line verification utilitzing multiple neural networks," *Optical Engineering*, vol. 36, pp. 3127–3133, 1997.

[35] P. P. A. Williams and L. Ronk, "Expert assissted, adaptive, and robust fusion architecture," *Optical Engineering*, vol. 37, pp. 378–390, 1998.

[36] J. Hansen, "Combining predictors, meta machine learning methodsand bias/variance ambiguity decomposition," *PhD thesis, University of Aarhus*, 2000.

[37] H. T. B. Duerr, W. Haettich and G. Winkler, "A combination of statistical and syntactical pattern recognition applied to classification of unconstrained handwritten numerals," *Pattern Recognition*, vol. 12, pp. 189–199, 1980.

[38] K. Fu, "Towards unification of syntactic and statistical pattern recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 5, pp. 200–205, 1983.

[39] P. Ahmed and C. Suen, "Computer recognition of totally unconstrained handwritten zip codes," *Intl. Journal of Pattern Recognition and Artificial Intelligence*, vol. 1, pp. 1–15, 1987.

[40] E. Mandler and J. Schurmann, "Combining the classification results of independent classifiers based on the dempster-shafer theory of evidence," *Pattern recognition and artificial intelligence*, pp. 381–393, 1988.

[41] L. Hansen and P. Salamon, "Neural network ensembles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 993–1001, 1990.

[42] A. K. L. Xu and C. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 22, pp. 418–435, 1992.

[43] T. M. R. L. C.Y. Suen, C. Nadal and L. Lam, "Recognition of totally unconstrained handwritten numerals based on theconcept of multiple experts," *In Intl. Workshop on Frontiers in Handwritting Recognition*, pp. 131–143, 1990.

[44] R. L. C. Nadal and C. Suen, "Complementary algorithms for the recognition of totally unconstrained handwritten numerals," *In 10 Intl. Conference on Pattern Recognition*, pp. 434–449, 1990.

[45] F. Alkoot and J. Kittler, "Multiple expert system design by combined feature selection and probability level fusion," *In FUSION 2000*, 2000.

[46] G. Giacinto, "Design of multiple classifier systems," *PhD thesis, University of Salerno*, 1998.

[47] R. Battiti and A. Colla, "Democracy in neural nets: voting schemes for classification," *Neural Networks*, vol. 7, pp. 691–707, 1994.

[48] L. Lam and C. Suen, "Application of majority voting to pattern recognition: An analysis of its behaviour and performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 553-:553, 1997.

[49] C. Ji and S. Ma, "Combinations of weak classifiers," *IEEE Transactions on Neural Networks*, vol. 8, no. 32-42, 1997.

[50] E. Alpaydin, "Multiple neural networks and weighted voting," *In Intl. Conf. Pattern Recognition*, pp. 29–32, 1992.

[51] L. Lam and C. Suen, "Optimal combinations of pattern classifiers," *Pattern Recognition Letters*, vol. 16, pp. 945–954, 1995.

[52] R. Schapire, "A brief introduction to boosting," *In: Proc. 16th Int. Joint Conf. AI*, 1999.

[53] R. Schapire, "The strength of weak learnability. machine learning,," *Machine Learning*, vol. 5, no. 2, 1990.

[54] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, 1996.

[55] L. Breiman, "Arcing classifiers," *The Annals of Statistics*, vol. 26, no. 3, 1998.

[56] F. Y. B. P. L. W. Schapire, E., "Boosting the margin: A new explanation for the effectiveness of voting methods," *The Annals of Statistics*, vol. 26, no. 5, 1998.

[57] G. Webb, "Multiboosting: A technique for combining boosting and wagging," *Machine Learning*, 2000.

[58] C. D. M. Chibelushi, "Adaptive classifier integration for robust pattern recognition," *IEEE Transactions on System, Man and Cybernetics*, vol. 29, December 1999.

[59] Y. Huang and C. Suen, "A method of combining multiple experts for the recognition of unconstrained handwritten numerals," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 90-94, 1995.

[60] K. K. H. Kang and J. Kim, "Optimal approximation of discrete probability distribution with kth-order dependency and its application to combining multiple classifiers," *Pattern Recognition Letters*, vol. 18, no. 515-523, 1997.

[61] K. K. H. Kang and J. Kim, "A framework for probablistic combination of multiple classifiers at an abstract level," *Engineering Applications of Artificial Intelligence*, vol. 10, no. 379-385, 1997.

[62] J. H. T.K. Ho and S. Srihari, "Decision combination in multiple classifier system," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, pp. 66–75, 1994.

[63] F. Alkoot and J. Kittler, "Experimental evaluation of expert fusion strategies," *Pattern Recognition Letters*, vol. 20, pp. 1361–1369, 1999.

[64] G. Rogova, "Combining the results of several neural network classifier," *Neural Networks*, vol. 7, pp. 777–781, 1994.

[65] D. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, pp. 241–260, 1992.

[66] C. Merz, "Using correspondence analysis to combine classifiers," *Machine Learning*, vol. 36, no. 33-58, 1999.

[67] M. Greencare, "Theory and application of correspondence analysis," *Academic Press, London*, 1984.

[68] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation and active learning.," *In G. Tesauro, D.S. Touretzky, and T.K Leen, editors, Advances in Neural information Processing Systems 7. Cambridge, Mass MIT Press*, 1995.

[69] K. L. Y.S. Huang and C. Suen, "The combination of multiple classifiers by neural network approach.," *International Journal of Pattern Recognition and ArtificialIntelligence*, vol. 9, pp. 579–597, 1995.

[70] G. Shafer, "A mathematical theory of evidence," *Princeton University Press*, 1976.

[71] I. B. N. Milisavljevic and M. Acheroy, "Characterization of mine detection sensors in terms of belief functions and their fusion, first results," *In 3rd Intl. Conf. on Information Fusion*, 2000.

[72] P. Smets, "The transferable belief model for quantified belief representation," *In D.M. Gabbay and P. Smets, editors,Handbook of defeasible reasoning and uncertainty*, vol. Kluwer, pp. 267–301, 1998.

[73] G. A. von Graevenitz, "About speaker recognition techology," *Bergdata Biometrics GmbH*, 2000.

[74] D.A.Reynolds, "A gaussian mixture modeling approach to text-independent speaker identification," *Ph.D Thesis, Georgia Inst. of Technology*, Sept 1992.

[75] R. D.A.Reynolds and M.J.T.Smith, "Pc-based tms320c30 implementation of the gaussian mixture model text-independent speaker recognition system," *Proc. Int. Conf. Signal Processing Appl., Technol.*, pp. 967–973, Nov 1992.

[76] G.McLachlan, *Mixture Models*. New York: Marcel Dekker, 1988.

[77] N.Z.Tishby, "On the application of mixture ar hidden markov models to text independent speaker recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 563–570, March 1991.

[78] A.N.Poritz, "Linear predictive hidden markov models and the speech signal," *Proc. IEEE ICASSP*, pp. 1291–1294, May 1982.

[79] A.E.Rosenberg, "Sub-word talker verification using hidden markov models," *IEEE ICASSP*, pp. 269–272, April 1990.

[80] D. Levinson, S.E.; Roe, "A perspective on speech recognition," *Communications Magazine, IEEE*, vol. 28, January 1990.

[81] M. Kohata, "Interpolation of lsp coefficients using recurrent neural networks," *Electronics Letters*, vol. 32, August 1996.

[82] W. J. Ephraim, Y. and L. Rabiner, "A linear predictive front-end processor for speech recognition in noisy environments," *IEEE Transactions ASSP*, 1987.

[83] B.S.Atal, "Effectivness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *JASA*, pp. 1304–1312, 1974.

[84] A. X.Haung and H.W.Hon, *Spoken Language Processing: A guide to theory, algorithm and system development.* Prentice Hall PTR, New Jersey,, 2001.

[85] B.C.J.Moore, *Frequency Analysis and Masking.* Academic Press, USA, 1995.

[86] B.C.J.Moore, "Information extractin and perceptual grouping in the auditory system," *Human and Machine Perception: Information Fusion*, 1997.

[87] I.-C. F.Bimbot and L.Mathan, "Second-order statistical measure for text-independent speaker identificatin," *Speech Communication*, vol. 17, pp. 177–192, 1995.

[88] P. J. F. K. I. Chang, K. W. Bowyer, "Face recognition using 2d and 3d facial data," *In Proc. of Workshop on Multimodal User Authentication, (Santa Barbara, CA)*, pp. 25–32, Dec. 2003.

[89] H. C. S. A. K. J. A. Kumar, D. C. M. Wong, "Personal verification using palmprint and hand geometry," *In Proc. of 4th Intl Conf. on Audio and Video based Biometric Person Authentication (AVBPA) (Guildford, UK)*, pp. 668–678, jun 2003.

[90] S. C. A. K. Jain, S. Prabhakar, "Combining multiple matchers for a high security fingerprint verification system," *Pattern Recognition Letters*, vol. 20, pp. 1371–1379, 1999.

[91] J. A.Ross, A..k.Jain, "A hybrid fingerprint matcher," *Pattern Recognition*, vol. 36, pp. 1661–1673, jul 2003.

[92] A. K. J. X. Lu, Y. Wang, "Combining classifiers for face recognition," *in Proc. IEEE Intl Conf. on Multimedia and Expo*, vol. 3, pp. 13–16, jul 2003.

[93] D. R.Brunelli, "Person identification using multiple cues," *IEEE Transactions on PAMI*, vol. 12, pp. 955–966, oct 1995.

[94] B. D. S. F. E.Bigun, J.Bigun, "Expert conciliation for multimodal person authentication systems using bayesian statistics," *in First International Conference on AVBPA, (Crans-Montana, Switzerland)*, pp. 291–300, march 1997.

[95] A. K. J. L. Hong, "Integrating faces and fingerprints for personal identification," *IEEE Transaction on PAMI*, vol. 20, pp. 1295–1307, dec 1998.

[96] U. D. R. W. Frischholz, "Bioid: A multimodal biometric identification system," *IEEE Computer,*, vol. 33, no. 2, pp. 64–68, 2000.

[97] B. Zhang and S. N.Srihari, "Class-wise multi-classifier combination based on dempster-shafer theory," *Seventh Internation Conference on Control, Automation, Robotics and Vision (ICARCV'02)*, pp. 698–703, Dec 2002.

Vitae

- Imran Naseem.
- Born in Karachi, Pakistan on February 20, 1979.
- Received Bachelor of Engineering (B.E) degree in Electrical Engineering from N.E.D University of Engineering and Technology, Karachi, Pakistan in 2002.
- Joined King Fahd University of Petroleum and Minerals in February 2003.
- Publications: "A Hybrid PCA-ANN Approach to Face recognition", Mohamed Deriche and Imran Naseem, in the 2nd IEEE GCC conference 2004.
- Email: inaseem@kfupm.edu.sa